2-1-2020

JUST A BIG MISUNDERSTANDING? BIAS AND BAYESIAN AFFECTIVE POLARIZATION

Daniel F. Stone
Bowdoin College

INTERNATIONAL ECONOMIC REVIEW
Vol. 61, No. 1, February 2020 DOI: 10.1111/iere.12421

JUST A BIG MISUNDERSTANDING? BIAS AND BAYESIAN AFFECTIVE POLARIZATION*

By Daniel F. Stone[1]

Bowdoin College, U.S.A.

I present a model of affective polarization—growth in hostility over time between two parties—via quasi-Bayesian inference. In the model, two agents repeatedly choose actions. Each choice is based on a balance of concerns for private interests and the social good. More weight is put on private interests when an agent's character is intrinsically more self-serving and when the other agent is believed to be more self-serving. Each agent Bayesian updates about the other's character, and dislikes the other more when she is perceived as more self-serving. I characterize the effects on growth in dislike of three biases: a prior bias against the other agent's character, the false consensus bias, and limited strategic thinking. Prior bias against the other's character remains constant or declines over time, and actions do not diverge. The other two biases cause actions to become more extreme over time and repeatedly be "worse" than expected, causing mutual growth in dislike, that is, affective polarization. The magnitude of dislike can become arbitrarily large—even when both players are arbitrarily "good" (unselfish). The results imply that seemingly irrelevant cognitive biases can be an important cause of the devolution of relationships, in politics and beyond, and that subtlety and unawareness of bias can be key factors driving the degree of polarization.

"Most quarrels amplify a misunderstanding."
- André Gide

"People are not so complicated. Relationships between people are complicated."
- Amos Tversky

1. INTRODUCTION

Why is it so hard to resolve disagreements? Persistent disagreement and bargaining impasses have been studied extensively by economists, but remain puzzling.[2] One factor usually excluded from economic analysis of these topics is interpersonal feelings. But disagreement often yields dislike, and vice versa. Feelings can in fact have first-order effects on outcomes, as differences of opinion and hostility sometimes build on one another, leading relationships to completely fall apart. We are all familiar with examples involving friends, family, colleagues, and business partners.[3]

*Manuscript received November 2017; revised November 2018.

[1] I thank Hülya Eraslan (the Associate Editor) and four anonymous referees for very helpful comments. I also thank Steven J. Miller, Dan Wood, Gaurav Sood, Dan Kahan, Andy Rudalevige, Roland Bénabou, Andrei Shleifer, and participants at talks at the 2017 AEA meetings, 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 suggestions and discussion. This article was written while I visited the University of Virginia's Department of Economics; I am grateful for their hospitality. Please address correspondence to: D. F. Stone, Department of Economics, Bowdoin College, Brunswick, ME 04011. E-mail: *dstone@bowdoin.edu*.

[2] See Williams (2017) and Loh and Phelan (2019) for recent models and discussion of the disagreement literature, which I also discuss briefly in Section 2. See Babcock and Loewenstein (1997) for discussion of work on bargaining impasses.

[3] A poignant example for behavioral economists is the falling-out between Daniel Kahneman and Amos Tversky described by Lewis (2016).

This phenomenon also seems to have occurred in U.S. politics over recent decades. Both hostility and partisan impasses over policy have increased substantially over this time (Snowe, 2013; Hetherington and Rudolph, 2015). Political scientists have dubbed the now well-documented growth in hostility "affective polarization" to distinguish it from polarization of underlying ideologies, or lack thereof (Iyengar et al., 2012; Lelkes, 2016). This term is fitting for the growth in hard feelings that regularly occurs in a wide range of repeated bilateral interactions.

A cynical explanation for affective polarization is that the hostility is warranted. Each side grows to understand, correctly, that the other deserves to be loathed. This explanation may have some truth to it but is certainly not completely satisfying, especially given how often the phenomenon seems to occur. In this article, I present a model of a more optimistic alternative: affective polarization driven by misunderstanding. I model inter-personal feelings by assuming they are based on cognitive beliefs, consistent with recent work from psychology and neuroscience.[4] I seek to complement other theories of hostility, such as "mindless" emotional reaction, social group identity, and motivated reasoning, by showing how hostility can also occur and grow due to Bayesian inference with biased priors ("quasi-Bayesian" inference; Benjamin, 2018).[5] These biases are relevant to a wide range of bilateral relationships, including those that do not involve social groups, and thus provide an explanation for the general tendency toward undue conflict in such relationships.

In the model, there are two agents, L and R (left and right). Each repeatedly chooses an action x based on private interests and tastes for the common good. Private interests are diametrically opposed and tastes may be aligned or differ in an unobserved way. If one interprets the agents as representative partisans, then it is natural to interpret x as an action reflecting larger or smaller government services and expenditures: L privately benefits, on average, from larger government, and R from smaller government, and L and R may have different tastes regarding the x that best serves society. Each agent balances her tastes for the common good and private interest in choosing x in each period based on her own "selfishness" parameter s_i, and beliefs about s_{-i}: i puts more weight on private interests when s_i is higher and, if there are reciprocity motives, which may be either intrinsic or strategic, when $E_i(s_{-i})$ is higher. Each agent i updates beliefs about s_{-i} after observing $-i$'s last choice of x plus noise.

Ex post affective polarization occurs when $E_L(s_R)$ and $E_R(s_L)$ grow due to high realizations of the s's. I provide two formal definitions of stronger forms of affective polarization that occur on average ex ante. The weaker definition requires that $E_i(s_{-i})$ tends to increase over time for each i even for average (as well as above average) realizations of the s_i's. The stronger definition requires both $E_L(s_R)$ and $E_R(s_L)$ to grow for *all* realizations of the s_L and s_R, that is, even when the agents are arbitrarily "good."

If beliefs converge toward truth, neither form of ex ante affective polarization occurs. In order to look for causes of such polarization, I examine the effects of three plausibly relevant biases in priors. I assume complete unawareness of these biases, that is, the agents' priors are technically misspecified, as they place probability zero on a feasible event (that the other agent is biased). This benchmark assumption makes the analysis tractable, and enhances the magnitude of some results, but does not drive their direction.

The first prior bias that I consider is direct bias against the other agent's character: overestimation of her s_i prior to the start of play. There is a large psychology literature on such "out-group" biases against other social groups (Ruffle and Sosis, 2006). It is possible that if the agents have a reciprocity motive, this bias could build on itself. This is not the case, however.

[4] From social psychology Haidt (2012) refers to "a prevalent but useless dichotomy between cognition and emotion," and from neuroscience, Pessoa (2008) says "parcelling the brain into cognitive and affective regions is inherently problematic, and ultimately untenable." For related work from political science see Mercer (2010), and from philosophy, Nussbaum (2003).

[5] See Mason (2018) for a summary of research on the importance of strengthened partisan identity, "tribalism," and motivated reasoning driving hostility in politics. New media (that we are excessively exposed to, and credulous of, negative views of the other side; see, e.g., Lelkes et al., 2015) is another factor relevant specifically to politics.

Even with full reciprocity (equal weight on s_i and $E_i(s_{-i})$), the bias does not grow over time, and with a smaller reciprocity motive, the bias declines over time. In order to see this, consider L's updating about s_R (the model is completely symmetric so R's updating is analogous). When L overestimates s_R, R acts less selfishly than *L* anticipates, causing the bias in $E_L(s_R)$ to decline. This process is tempered if R acts more selfishly due to reciprocation of an upward-biased $E_R(s_L)$. But the process never exacerbates bias because *R*'s bias cannot make R act more selfishly than L anticipates, relative to true s_R.

The next bias that I examine is the false consensus effect: overestimation of similarity of tastes. The false consensus bias is well established in social psychology, but this bias has only been studied to a limited extent in behavioral economics. I assume that tastes for the action that best serves the social good are correlated with private interests. For example, L is likely to truly believe the larger government action is better for society, though this action also privately benefits L, and R's tastes and benefits are similarly correlated. The false consensus bias causes the agents to underestimate true differences in tastes. I show that this bias causes the weaker form of ex ante affective polarization, even in the absence of a reciprocity motive. By underestimating the differences in tastes, each agent consequently overattributes actions driven partly by such tastes to selfish motives. However, without reciprocity motives, the degree of the stronger form of affective polarization and long-run bias is very limited.

If there is a reciprocity motive, the stronger form can occur for arbitrarily small initial bias. Moreover, for arbitrarily small initial bias and unselfishness of both agents, there exist parameter values such that the magnitude of affective polarization can be arbitrarily large. The intuition is as follows. When L acts "selflessly" in period 1, L expects R to update beliefs about s_L downward, and R indeed does this. But R does not do this to the extent that L expects, due to the false consensus bias (for all values of s_R and s_L). Thus, the reciprocation effect causes R's action in period 2 to make R's character in period 2 look worse to L than R realizes. This makes L reciprocate excessively in period 3. The cycle continues and compounds, whether the initial bias and true s_L and s_R are large or small. A race is then on between the increase in precision of beliefs over time, causing belief revisions to decline, and exacerbation of bad actions due to exacerbated misunderstanding. Although the basic model assumes myopic agents, I show in an extension that results are robust to allowing the agents to have foresight, creating a strategic motive to moderate actions.

I next consider a jointly chosen policy in each period, creating strategic incentives even for myopic agents. If the agents were sophisticated strategic thinkers, they would choose opposing extreme private contributions to the joint policy to counterbalance one another. In this case, the agents do not consider such private actions to be signs of bad character, just sound strategy. Thus, adding such strategic concerns to the model does not inherently cause affective polarization. However, there is also a new parameter to consider for this case: the degree of strategic thinking. I discuss the plausibility of limited strategic thinking in politics, and show that if both agents are equally limited (Level k) strategic thinkers, they underestimate the strategic component of their opponent's actions, and overinfer the extent to which these actions reflect self-serving motives, causing dislike to grow. The jointly chosen action creates an endogenous reciprocity motive. Consequently, again growing dislike leads to growing extremism of actions. Again, the process can compound to any extent, and the stronger form of affective polarization can occur, for some parameters. Also, again, these results occur whether or not the agents are forward-looking.

In summary, I take an economic framework (payoff maximization and Bayesian inference) to the analysis of dislike, assuming it results from inferences about self-serving character. I consider the effects of four additional psychological elements: three biases (false consensus, level k thinking, out-group), and a reciprocity motive. I show the two biases that are not directly related to character, false consensus and level k thinking, cause Bayesian affective polarization, which can snowball to a striking degree over time due to reciprocity. The bias directly related to character (out-group bias) does not cause character impressions to worsen. These results imply that biases indirectly related to character may be more pernicious, since they are more likely to cause actions to be consistently "worse" than expected.

More broadly, the analysis implies that in a range of bilateral settings involving private and shared interests, (1) feelings affect choices and feelings are driven by beliefs, (2) some standard cognitive errors make the other agent's actions appear excessively self-serving, leading to biased, negative inferences and hard feelings, (3) such biased beliefs cause misleadingly self-serving actions, (4) misleading actions cause even more biased beliefs, and so this process can build on itself. This general point, about how bias-driven hostility tends to snowball, could apply to other biases beyond those studied here.[6] In the final section of this article, I elaborate on issues with the analysis and directions for future work.

2. ADDITIONAL RELATED LITERATURE

To my knowledge, this article is the first from economics to address affective polarization, and the first across disciplines to discuss the relationship between "unmotivated" biases (those not driven by motivated reasoning) and political affective polarization.[7] The economics literatures most closely related to this article are those on political extremism and polarization, on disagreement more generally, on misspecified priors, and on ethnic conflict.

Most economics papers on political polarization do not incorporate bias. One basic question of interest is what causes extremism. One paper that incorporates cognitive bias in answering this question is Ortoleva and Snowberg (2015): They provide a theoretical argument, and empirical evidence, that bias (overprecision) causes ideological extremism. A related paper is Blomberg and Harrington (2000); they show that if priors are heterogeneous, a positive correlation of political extremism and rigidity (precision) of beliefs may arise simply due to Bayesian updating. However, neither of these papers study dynamic growth in extremism, or partisan hostility.

As noted in the introduction, persistent disagreement has been considered puzzling to economists, at least since Aumann (1976) showed rational agents should not agree to disagree, and many later papers have proposed models to try to explain this empirically common occurrence. For example, Sethi and Yildiz (2012) show that if priors are heterogeneous, unobservability of priors exacerbates disagreement, as does segregation of social groups with different priors. Baliga et al. (2013) show that disagreement can increase (i.e., beliefs can polarize) after observing a common signal if agents are ambiguity averse. Andreoni and Mylovanov (2012) provide a model, and experimental evidence, showing that heterogeneous priors on one dimension of the state of the world cause diverging responses to new information on another dimension. An important similarity between our models is that different beliefs about one dimension (in my model, tastes or level of strategic thinking) is the cause of increasing disagreement on the other dimension (character). Papers on related identification problems (data that are lower dimensional than the true model) leading to persistent or growing disagreement include Acemoglu et al. (2016), Piketty (1995), Benoît and Dubra (2019), and Loh and Phelan (2019). These papers are quite distinct from mine in two ways: (1) They do not focus on cognitive bias as the root cause of disagreement; (2) they do not address interpersonal feelings, or growing dislike in particular.

Fudenberg et al. (2017) provide a helpful discussion of the literature on misspecified models, that is, those with agents who place prior probability zero on a feasible event. In their analysis, they study the long-run effects on beliefs and actions of the degree to which an agent is forward-

[6] Consistent with this broader view, in a companion paper I show that a related, but distinct, bias empirically predicts partisan hostility, conditional on ideology and partisanship (Stone, 2019).

[7] There is a strand of literature in psychology on how related cognitive biases can cause problems in interpersonal relationships such as marriages (Ross et al., 1979; Bradbury et al., 1996). I am not aware of any literature on how the specific biases that I study in this article cause relationship problems, or on the relation between these (or any unmotivated) biases and Bayesian inference.

looking. Their focus is on whether or not beliefs converge when the true state is not in the support of the agent's prior. This differs from my article, in which the true state of interest (s_{-i}) is always in the support of each agent's prior, and the main question is how much higher than the true value do expectations grow due to a prior misspecified in another dimension. Similarly, Esponda and Pouzo (2016) is only indirectly related, as they develop a solution concept for games with players who may have misspecified models. Heidhues et al. (2018) obtain a result more similar to mine, showing how unawareness of bias can result in a positive-feedback process in which the agent's actions lead to increasingly misleading signals, and consequently increasingly biased posterior beliefs. The setting I study is quite different, but our papers are complementary, in that, they focus on the effects of a bias that I do not study, overconfidence, which supports the idea that a variety of biases can have similar snowball-type effects. The generality of unboundedness of long-run bias in my model (that it can occur for any realization of the s's) is a unique result, to my knowledge, compared to papers in this literature and the disagreement literature. This result is driven by the reciprocity motive that is typically not appropriate for the contexts studied in those literatures.

There are numerous economics papers on group-based hostility. Glaeser (2005) models hatred toward 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 et al. (2016) is more similar to my article 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) 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 bilateral setting in which the two agents have both conflicting and shared interests. The model has the structure of an infinitely repeated game (it is convenient to use game theoretic terminology even for part of the analysis that is non-game theoretic). In Subsection 3.1, I describe the stage game set-up. In Subsection 3.2, I describe how the players update beliefs about each other across stages, and how actions depend on beliefs in each stage. In Subsection 3.3, I define two types of affective polarization that may occur. I include discussion of the model assumptions throughout the section.

3.1. *Players, Actions, and Payoffs.* There are two players, L (left) and R (right). The stage game payoff for each player i is

$$u_i(x_i, x_{-i}) = u_i^p(x_i) + u_i^p(x_{-i}) + \alpha_i(u^s(x_i; \tau_i) + u^s(x_{-i}; \tau_i)). \tag{1}$$

Stage subscripts are omitted for now. Player i chooses $x_i \in \mathbb{R}$ and receives direct, private payoffs from the $u_i^p(.)$'s, and subjective social payoffs from the $u^s(.)$'s. These are subjective because they depend on i's taste parameter $\tau_i \in \mathbb{R}$, but are not subscripted by i because the functional form is the same for both players. The prosociality of i—the extent to which i is willing to trade off private for social gains—is represented by $\alpha_i > 0$, 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 $i = R$ and $u_i^p(x) = -x$ for $i = L$ for the private payoffs. Thus,

$$u_L(x_L, x_R) = -x_L - x_R - \alpha_L\big((x_L - \tau_L)^2 + (x_R - \tau_L)^2\big);$$
$$u_R(x_R, x_L) = x_L + x_R - \alpha_R\big((x_L - \tau_R)^2 + (x_R - \tau_R)^2\big).$$

Each player makes a unilateral choice that has both a private and social payoff for both players. (In Section 6 the action is chosen jointly.) R privately benefits from both x_L and x_R being larger, but thinks "society" is best off when both are as close as possible to τ_R. L faces an analogous trade-off, privately preferring smaller x_i's. Individuals in bilateral relationships typically face such trade-offs between private and social benefits when making decisions. Note that the players' actions are not combined in the social payoff function. This can be interpreted as resulting from the actions relating to different issues, or occurring in different contexts. That is, the players may take actions in distinct settings, but each action has a social consequence.[8]

The prosociality weight α_i, is equal to $(1/2)(\frac{1}{s_i+rE_i(s_{-i})})$, with $s_i \in (0, \infty)$ for each i. The parameter s_i can be interpreted as i's "selfishness": as s_i increases, i puts relatively more weight on u_i^p and less on u_i^s. The term $rE_i(s_{-i})$ represents "reciprocal selfishness": if $r > 0$, i puts more weight on private interests, and less on social interests, as $E_i(s_{-i})$ grows.[9] It is natural to assume that $r \in [0, 1]$.[10] The coefficient of 1/2 just simplifies algebra. Thus, $\alpha_i \in (0, \infty)$, and α_i approaches zero as either s_i or $E_i(s_{-i})$ approaches ∞, and α_i approaches ∞ when both s_i and $E_i(s_{-i})$ approach zero (if $r > 0$).

For much of the analysis, I assume that the players are not forward-looking, and in each period, each agent simply maximizes subjective expected utility for the stage game. This assumption is made for tractability and is supported by the large literature on limited strategic foresight, see, for example, Jéheil (1995). I also explore a variant of the model in which the agents have some foresight.

The agents can be interpreted as representative politicians from the competing parties in the United States, with a more leftist (negative) x corresponding to more expansive government, which provides more direct private benefits to Democrats, on average, and a larger x corresponding to "smaller" government, which tends to provide private benefits to Republicans. Politicians and parties also often take unilateral actions, such as executive actions (which have increased over recent decades; see, e.g., Howell, 2003; Lowande, 2014; Belco and Rottinghaus, 2017), legislative proposals/amendments, and media statements on policy positions or party platforms.[11] In fact, some acts of major legislation passed in the last decade have lacked any bipartisan support (Obama's health care reform and Trump's tax reform), and other major bills have had extremely low bipartisan support (e.g., Obama's stimulus package), and thus have essentially been actions taken by one party or the other. These actions of course still affect future beliefs and actions taken by the other party and consequently still involve strategic concerns. The model modification in Section 5.3 addresses this issue (and again, Section 6 addresses simultaneous strategic play, with a jointly determined action).

A nonpolitical interpretation of the agents is that they are spouses, and have different tastes regarding how to allocate some resource, such as time. For example, suppose one is a

[8] The social payoff of an action 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 her own preferences).

[9] The payoff for i is a direct function of $E_i(s_{-i})$, rather than s_{-i}, to make the analysis more tractable.

[10] See, for example, Levine (1998) for an example of a model that incorporates preferences for reciprocity in a related way (with regard to others' types and not just their actions). See, for example, Batson and Powell (2003) for discussion of the (very large) broader literature on the importance of reciprocity for prosocial behavior. One could also think of r as a spitefulness parameter, since higher values cause the player to act in a way that makes the other player worse off, but I refer to the parameter as reciprocity since its effect depends directly on the other player's type.

[11] For discussion of the controversy and role in growing political anger due to executive actions taken in the most recent years, see Rudalevige (2018).

"workaholic" that in general privately prefers to work as much as possible. The household overall benefits from this to a point, but the two spouses disagree about the socially optimal level of work (which may affect both spouses and other members of the household), and weight the trade-off between private and social interests in different ways.

3.2. *Beliefs, Learning, and Choices.* There is common knowledge of the payoff functional forms, and each agent i knows her own s_i but does not observe s_{-i}. I assume that there is a noise term added to $x_{i,t}$ in each stage game period t, $\epsilon^i_t \sim N(0, \sigma^2_\epsilon)$, and that $\hat{x}_{i,t} = x_{i,t} + \epsilon^i_t$ is observed by both players in each period (though i also knows $x_{i,t}$). Perhaps there are additional exogenous factors affecting each player's action in each period, or there is noise in how the action is publicly perceived. The addition of noise prevents the players from believing that they pinpoint each other's parameters with limited data, which could either stop the updating process (and thus stop affective polarization) or lead to the perceived occurrence of zero-probability events.

The true distribution of s_i, $i \in \{L, R\}$, is left-truncated normal, with lower bound zero, $\mu_s > 0$, and σ^2_s. I simplify the analysis by approximating this distribution with the non-truncated version for the analysis of belief updating. This approximation is highly accurate when μ_s is relatively far from zero, making the truncated and nontruncated distributions essentially equivalent. In order to be consistent with this approximation, in the numerical examples I assume that μ_s is at least four standard deviations from zero, implying $Pr(s_i > 0) > 0.9999$.[12] If τ_{-i} is unobserved, then for each i, $\tau_{-i} \sim N(\mu_{\tau^i_{-i}}, \sigma^2_\tau)$.

In order to condense notation, let $s^{i,t}_{-i}$ 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 $s^{i,-i,t}_i$ denote the second-order expectation $E_{i,t}(E_{-i,t}(s_i))$ (i's expectation at the end of period t of $-i$'s expectation at the end of t of s_i), and let the third-order expectation, and the expectations for τ, be defined analogously.

Using this notation, given stage–game payoff maximization, the period t choice functions are:

$$\begin{aligned} x^*_{L,t} &= -(s_L + r s^{L,t-1}_R) + \tau_L, \\ x^*_{R,t} &= s_R + r s^{R,t-1}_L + \tau_R. \end{aligned} \tag{2}$$

The extremism of i's action is increasing in s_i, in i's expectation of s_{-i} if $r > 0$, and in the degree to which τ_i is aligned with i's private interests, for $i = L$ and $i = R$.

Before proceeding, note that if i thinks she knows $-i$'s priors, and there is common knowledge about Bayesian updating, then $Var_{i,t}(s^{-i,t}_i) = 0$ for all i and t. This is because i observes all information that $-i$ observes through period t. Consequently, i is certain that she knows $-i$'s updated beliefs about i in period t. Similarly, if i thinks she knows $-i$'s beliefs about i's priors, then $Var_{i,t}(s^{-i,i,t}_{-i}) = 0$. Moreover, if i thinks that $-i$ holds correct beliefs about i's priors, then third-order beliefs are the same as first-order beliefs: i thinks $-i$ holds correct beliefs about i's priors and information, and thus also about i's updated beliefs, so $s^{i,-i,i,t}_{-i} = s^{i,t}_{-i}$ for all i and t. These points will be used extensively in the subsequent analysis.

3.3. *Affective Polarization.* I assume that agent i's dislike of $-i$ is an increasing function of $s^{i,t}_{-i}$. It is intuitive and supported by research that individuals form inferences about the moral character of others based on their actions (Uhlmann et al., 2015), and that a willingness to sacrifice self-interest for the social good is widely considered to be fundamental to morality and quality of character (Greene, 2014; Curry et al., 2018; Moshagen et al., 2018), which can in turn

[12] This assumption does imply that values of s_i very close to zero are highly improbable, but this seems very reasonable; moreover, the approximation becomes more accurate as beliefs about s_{-i} move away from zero, that is, as dislike grows (the focus of the analysis).

drive hostility (Haidt, 2012; Ryan, 2014; Garrett and Bankert, 2018).[13] 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 1. "Expected affective polarization" occurs iff: $E_0(s_R^{L,t})$ and $E_0(s_L^{R,t})$ are increasing in t.

Definition 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.

As noted in the introduction, both definitions require ex ante growth in dislike, and do not just refer to particular ex post realizations of the s's. "Expected affective polarization" occurs when both agents, on average over the realizations of the s's, grow to dislike each other more over time. "Strong affective polarization" occurs for all realizations of the s parameters. I also characterize the probability limits of expectations of s_{-i} to the extent possible. Clearly as these grow further above the prior mean, this also indicates affective polarization. I do not conduct any welfare analysis, since I do not model a true social welfare function, but since both forms of affective polarization imply growing inaccuracy in beliefs, they both imply a decline in the quality of choices and payoffs.

4. out-group bias

The first bias that I consider is a prior bias against the opposition side (the analysis of this case subsumes the case of no bias). As noted above, there is a substantial literature in social psychology on such an "out-group" bias (Tajfel, 1982). I model this by assuming that for each $i \in \{L, R\}$:

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

I also assume complete unawareness of this bias—that is, that each player i is unaware of both her own and $-i$'s bias, incorrectly believing that $-i$ knows i's prior. It is possible that affective polarization results from a small initial bias building on itself, especially with such unawareness of bias and myopic agents, but I show this is not the case. This also demonstrates that affective polarization is not "built into" the model. In order to simplify the analysis in this section, I assume that there is common knowledge that $\tau_L = \tau_R = 0$; I show in Appendix A.3 that this assumption is without loss of generality (results are qualitatively the same when each agent i is uncertain about τ_{-i}).

Consider belief updating by L about s_R (R's updating is symmetric). Since $\hat{x}_{i,t} = x^*_{i,t} + \epsilon^i_t$, the new information L observes in period t is $\hat{x}_{R,t} = s_R + rs_L^{R,t-1} + \epsilon^R_t$. L has no uncertainty about $s_L^{R,t-1}$, given unawareness of bias, as noted at the end of Subsection 3.2. Thus, $Var_L(\hat{x}_{R,t}|s_R) = \sigma^2_\epsilon$, and L updates her expectation of s_R in the standard way given a normal prior and normal signal, as a weighted average of the prior mean and the signal, $\hat{x}_{R,t}$, adjusted so that its mean is equal to the parameter value, s_R (by subtracting off $rs_L^{L,R,t-1}$):

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

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

[13] The affective polarization literature focuses on feelings of voters, and not politicians. In order to connect the representative politician interpretation of the agent to voters, one could either assume: (1) the politician has the same preferences and beliefs as her party's median voter, or (2) for political reasons the politician maximizes her party's median voter's payoff, given that voter's beliefs.

R *thinks* that L updates by:

$$\begin{aligned} s_R^{R,L,t} &= \lambda_t s_R^{R,L,t-1} + (1-\lambda_t)(\hat{x}_{R,t} - r s_L^{R,L,R,t-1}) \\ &= \lambda_t s_R^{R,L,t-1} + (1-\lambda_t)\Big((s_R + r s_L^{R,t-1} + \epsilon_t^R) - r s_L^{R,t-1}\Big) \\ &= \lambda_t s_R^{R,L,t-1} + (1-\lambda_t)(s_R + \epsilon_t^R). \end{aligned} \tag{4}$$

There are two points worth noting here. First, both $s_R^{L,t}$ and $s_R^{R,L,t}$ 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 article. Second, the second line uses the fact that third-order beliefs equal first-order beliefs as also noted at the end of Subsection 3.2: $s_L^{R,L,R,t-1} = s_L^{R,t-1}$. The following result is immediate.

Lemma 1. *Second-order expectations converge to truth:* $s_i^{i,-i,t}$ *converges in probability to* s_i, *for* $i \in \{L, R\}$.

The updating procedure for $s_i^{i,-i,t}$, stated in (4), is standard, and the signal is unbiased (its mean is the true value of the parameter, s_R). In this case, it is well known that even if the initial prior mean is biased, the posterior mean converges to truth.[14] Since R thinks the players have common priors and new information, R thinks that L knows how R updates her beliefs about s_L. Thus, R thinks L uses the correct adjusted signal $(\hat{x}_{R,t} - r s_L^{R,t-1})$ when in reality L's adjusted signal is incorrect.

Second-order expectations not only converge to truth, but are unbiased in every period in the sense that they reflect the Bayesian expectations from the correct prior given available information. Thus, the difference between first-order and second-order expectations can be viewed as the bias in the first-order expectation each period. This difference (each side of the first line of (4) subtracted from the corresponding side of (3)) yields:

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

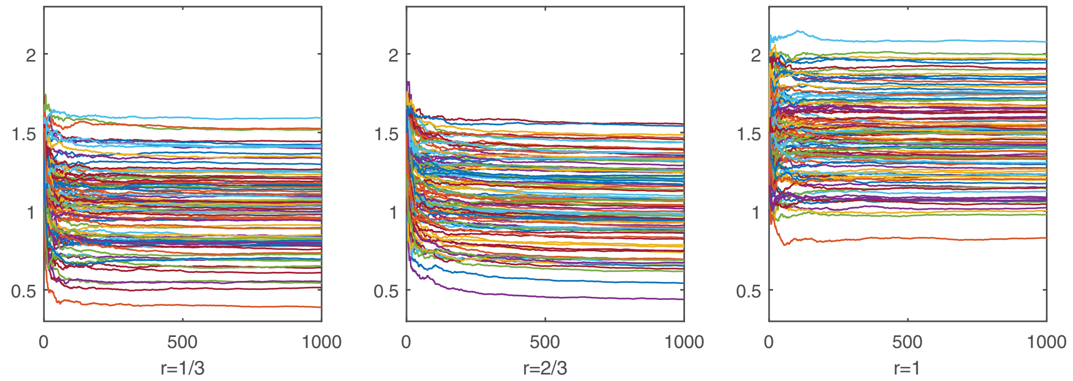
The difference is deterministic. It is driven only by the initial bias in priors, not random shocks to the $\hat{x}$'s. This is because these shocks affect first- and second-order beliefs in the same way. Furthermore, we can obtain a simple closed-form expression for this difference in beliefs by iterating forward. For $t = 1$:

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

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

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

[14] See, for example, Bullock (2009).



NOTES: Vertical axis $= E_{i,t}(s_{-i}) = s_{-i}^{i,t}$ versus time on horizontal axis. One hundred simulations per set of variance parameter values; prior mean $s_i = \mu_s = 1$ and prior bias against $-i = b = 0.5$ (and thus $(s_{-i}^{i,0} = 1.5)$); $\sigma_s^2 = \sigma_\epsilon^2 = 0.25$, for all simulations. $r =$ weight on $s_{-i}^{i,t}$ in i's prosociality parameter, α_i.

FIGURE 1

SIMULATED EVOLUTIONS OF i'S EXPECTATIONS OF $-i$'S SELFISHNESS WITH INITIAL OUT-GROUP BIAS FOR DIFFERENT VALUES OF RECIPROCITY CONCERNS [COLOR FIGURE CAN BE VIEWED AT WILEYONLINELIBRARY.COM]

The pattern, and symmetry, continue, so

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

This implies the following results (details for these and other results are provided in the appendix as necessary).

PROPOSITION 1. *With out-group bias, common knowledge of tastes and* $\tau_L = \tau_R = 0$:

1. *Neither form of affective polarization occurs.*
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. $\lim_{t\to\infty} E(x_{R,t}|s_L, s_R) = s_L + s_R + b = \lim_{t\to\infty} E(-x_{L,t}|s_L, s_R)$ *if* $r = 1$.

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 $(s_L + s_R + b)$. This action is stable because it is equal to what is expected. Once $s_R^{L,t}$ has become "close" to $s_R + b$ (and vice versa), since L's belief about s_R is biased upward by b, L's belief about next period's $x_{R,t}$ is also biased upward by b. But L's belief about R's belief about s_L is biased downward by b. This bias exactly cancels the upward bias, causing L's belief about next period's $x_{R,t}$ to be unbiased. L is therefore not "surprised" by the $\hat{x}_{R,t}$ observed on average, and so L does not revise beliefs about s_R on average. Note that if the agents had different prior biases, the biases would converge to the mean over time (so the smaller bias would grow but to a limited extent, and the larger one would decline).

If $r < 1$, then each player acts less selfishly than the other expects (on average). This causes i's beliefs about s_{-i} 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.

Figure 1 presents 100 simulations of the evolution of first-order beliefs for $t = 0 - 1000$, 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. In addition to showing that bias fully dissipates and beliefs converge to truth even for fairly high r ($r = 2/3$), this figure is also useful as a contrast to analogous figures presented later in the article.

5. FALSE CONSENSUS BIAS

The false consensus bias is, again, the tendency to overestimate the similarity between ourselves and others in some dimension—the bias toward thinking there is "consensus" in these things.[15] An extreme form of the false consensus bias occurs when an individual assumes that her preference is universally and objectively correct, when the issue might be a matter of taste or other factors. For example, a stickler about punctuality might think that it is "simply wrong" to ever be late, whereas others with different social norms or preferences might have no problem with a certain degree of lateness most of the time.

This application of the false consensus bias is novel but similar ideas have been expressed before, in particular by Haidt (2012), a very well-known work in political psychology. Haidt does not refer explicitly to the false consensus bias, but implies that this bias is an important cause of political discord, arguing that moral–political values are like tastes, and not objective truths, but that voters fail to understand this distinction.[16] However, Haidt does not, to my knowledge, specify a cognitive or inference-based mechanism for why this misperception causes bad blood. The model of this article helps to fill this gap.

For the analysis of this bias, it is natural to assume that L and R are truly different in tastes, at least on average, so I assume $\mu_{\tau_R} > \mu_{\tau_L}$. L likely has a true taste for a more leftist policy due to selection (e.g., partisans select into their party due to true political tastes) and/or motivated reasoning (the desire to believe that what is best for oneself is also best for society overall). I formalize the false consensus bias as follows:

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

Each player is again assumed to know the correct distribution for her own type, and be unaware of both her own bias and the other player's bias.

5.1. *No Reciprocity.*

First consider the case of no reciprocity ($r = 0$). In this case, $\hat{x}_{R,t} = s_R + \tau_R + \epsilon_1^R$. L's updated expectation of s_R after the first period is:

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

Taking objective expectations: $E_0(s_R^{L,1}) = \mu_s + (1 - \lambda_1)b$ since $E_0(\tau_R - \tau_R^{L,0}) = b$. Simply observing one action by the other player causes dislike of that player to increase, on average. Since the bias is symmetric, we have already obtained "expected affective polarization." Each player underestimates how different the other's tastes are from her own, so each likely observes the other taking an action that appears more self-serving than it really is.

In order 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

[15] See Ross et al. (1977) for early work from psychology and Butler et al. (2015) for a recent application from economics. Related biases include the curse of knowledge, egocentrism, and mindblindness (Pinker, 2015). The false consensus bias is more likely to apply to what might be called horizontally differentiated characteristics, such as certain types of tastes; for vertically differentiated characteristics, such as ability, overconfidence may be more likely to apply.

[16] From economics, McMurray (2017) makes a similar argument, that voters perceive of politics "through the lens of truth and error" rather than a contest of heterogeneous preferences. McMurray says that economists have traditionally assumed heterogeneous preferences to be the more accurate modeling assumption, and implies that voters underestimate the importance of heterogeneous preferences.

full set of signals: $\bar{x}_{R,t} = \frac{1}{t}(\hat{x}_{R,1} + \hat{x}_{R,2} + \cdots + \hat{x}_{R,t})$. This sample mean naturally accounts for the way in which the observations are correlated (their dependence on fixed s_R and τ_R), with $E_{L,0}(\bar{x}_{R,t}|s_R) = s_R + \tau_R^{L,0}$, and $Var_{L,0}(\bar{x}_{R,t}|s_R) = Var_0(\bar{x}_{R,t}|s_R) = \sigma^2_{\tau_R} + \sigma^2_\epsilon/t$. The observations are i.i.d. given $r = 0$, so no information is lost in combining them this way. L then updates her expectation (from the period 0 prior, after observing the first t signals) to:

$$\begin{aligned} s_R^{L,t} &= \lambda_t s_R^{L,0} + (1-\lambda_t)\Big(\bar{x}_{R,t} - \tau_R^{L,0}\Big) \\ &= \lambda_t \mu_s + (1-\lambda_t)\left(s_R + \tau_R - \tau_R^{L,0} + \frac{1}{t}\sum_{i=1}^{t} \epsilon_i^R\right), \end{aligned} \tag{10}$$

with $\lambda_t = \frac{Var_0(\bar{x}_{R,t}|s_R)}{\sigma_s^2 + Var_0(\bar{x}_{R,t}|s_R)}$. It is then clear that $E_0(s_R^{L,t}|s_R, \tau_R) = \lambda_t\mu_s + (1-\lambda_t)(s_R + \tau_R - \tau_R^{L,0})$, and $E_0(s_R^{L,t}|s_R) = \lambda_t\mu_s + (1-\lambda_t)(s_R + b)$, with λ_t decreasing in t. The following results are implied.

Proposition 2. *With uncertainty in tastes, false consensus bias b, and $r = 0$:*

1. *expected affective polarization occurs for all $b > 0$, for all t;*
2. *strong affective polarization never occurs if $b < \mu_s$;*
3. *plim $s_R^{L,t} = \frac{\sigma_\tau^2}{\sigma_s^2+\sigma_\tau^2}\mu_s + \frac{\sigma_s^2}{\sigma_s^2+\sigma_\tau^2}(s_R + \tau_R - \mu_{\tau_R} + b)$, and plim $s_L^{R,t}$ is analogous. The marginal effect of b on plim $s_{-i}^{i,t}$ is increasing in σ_s^2 and decreasing in σ_τ^2. The variance of i's belief about s_{-i} does not converge to zero.*

Consider two spouses, one of whom is responsible for preparing meals, and has a relatively strong preference for Italian food. If the other spouse knows of this relative preference, but underestimates its magnitude, then when the meal-preparer chooses to make Italian food this choice will appear more self-serving than it really is. This trend grows over time, as more weight is put on new information and less on the prior. This is why expected affective polarization occurs and persists. Strong affective polarization never occurs if b is not too large since then small s_R will cause L's beliefs about s_R to decline. The degree to which first-order expectations increase is limited by the magnitude of b, as implied by the third result (this implies that, for instance, if s_R and τ_R equal their mean values, then $s_R^{L,t}$ is bounded below $s_R + b$). Note that results reverse (dislike declines over time) if the direction of the bias is reversed.

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 σ_τ^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).

Although this affective polarization result is straightforward to derive and is similar to results found in other contexts as noted in Section 2, it is not necessarily intuitive—we might assume that a Bayesian updater with unbiased signals would learn the truth over time, or move toward truth, even with a biased prior, since again that is the case with just one unknown parameter. With two unknowns, bias in a prior for one causes a biased posterior for the other. However, in this case, there is no polarization of actions, and strong affective polarization is very limited, as is the degree of long-run hostility. I next incorporate reciprocity to see how this affects the results.

5.2. *Reciprocity.* Consider the benchmark case of $r = 1$. In $t = 1$, L now subtracts off both $s_L^{L,R,0}$ and $\tau_R^{L,0}$ from $\hat{x}_{R,1}$ to obtain a signal with expected value s_R, so L's updated expectation for s_R is:

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

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

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

since R is unaware of L's biased prior about τ_R and thus

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

This value is the same as that of the case of $r = 0$, and is again symmetric for beliefs about s_L.

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

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

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

So, in period 2, L updates by:

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

The first-order expectation is now further biased upward as compared to after $t = 1$, due to influence from both the initial false consensus bias and the bias this causes in R's expectation of s_L at the end of $t = 1$, $(1-\lambda_1)b$. That is, the false consensus bias causes R to overestimate s_L after $t = 1$, which causes R to take a more extreme action in $t = 2$ due to reciprocity, which in turn causes L to overestimate s_R even more after $t = 2$ as compared to after $t = 1$.

Let $b_t := s_{-i}^{i,t} - s_{-i}^{-i,i,t}$. This difference is again deterministic because both players condition on the ϵ's in the same way, and again symmetric. L's beliefs about s_R in any period t can be written as follows:

$$s_R^{L,t} = \lambda_t\mu_s + (1-\lambda_t)\left(s_R + \frac{1}{t}\sum_{i=1}^{t}\epsilon_i^R\right) + (1-\lambda_t)\left(\frac{1}{t}\sum_{i=1}^{t-1} b_i + \tau_R - \tau_R^{L,0}\right). \tag{16}$$

Since the first two terms of this expression are the same for $s_R^{R,L,t}$, we can subtract this second-order expectation from the first-order one to get

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

In the Appendix, I show that if $(1-\lambda_t) = 1$ for all t, then b_t is equal to b times the harmonic sum $(1 + 1/2 + 1/3 \ldots)$, which diverges. This implies that there exist values of the σ's such that b_t, and thus $s_R^{L,t}$, can become arbitrarily large—for any $b > 0$, $s_R > 0$, and $s_L > 0$! The next proposition follows from this analysis.

PROPOSITION 3. *With uncertainty in tastes, false consensus bias b, and $r = 1$:*
1. expected affective polarization occurs for all values of the parameters, for all t;
2. plim $s_R^{L,t} = \frac{\tau_s^2}{\sigma_s^2+\sigma_{\tau_R}^2}\mu_s + \frac{\sigma_s^2}{\sigma_s^2+\sigma_{\tau_R}^2}(s_R + \tau_R - \mu_{\tau_R} + b) + \lim_{t\to\infty} b_t$, 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 σ_τ^2 and σ_ϵ^2, for any b, s_L, s_R, τ_L, and τ_R;
3. there exists t' such that strong affective 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.

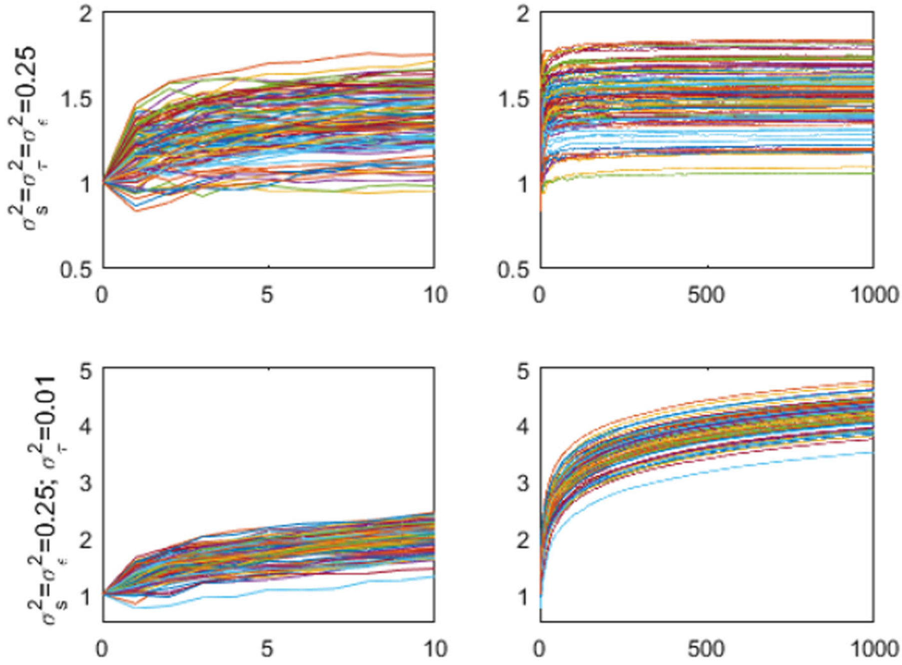
The proposition formalizes the "arbitrarily large" $s_{-i}^{i,t}$ (for any b and for any realizations of the s parameters) result, and also asserts that strong polarization occurs (eventually) for arbitrarily small initial bias, for similar values of the other parameters (the σ^2's) implying high weight is placed on new information. Again, strong affective polarization means increasing $E_0(s_{-i}^{i,t}|s_i, s_{-i})$ for each i, for all realizations of the s's.

These results occur because b_t increases over time for all realizations of the s's and any $b > 0$. 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 overestimates 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, whereas b_t always grows.

Figure 2 presents simulated updated first-order expectations in the relatively short run (left graphs) and long run (right graphs). I present the paths of expectations for a sample of draws from the parameter distributions to illustrate the distribution of updating dynamics. The figure shows that strong polarization occurs for reasonable values of the σ^2's, and typically starts quite early (low t). In the left graphs, although beliefs decline for some simulations in the first few periods, likely when s_{-i} is indeed low, beliefs quickly and steadily begin to rise again for *all* simulations. The bottom right graph shows how expectations can continue to grow over an extended period of time, and can grow relatively very large (they approximately quadruple on average) when priors for the τ's are very precise.

The strength of these results is driven by the strength of the unawareness of bias assumption, implying that the players never update toward bias being the culprit for the growing extremism in actions. This assumption seems relatively 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 relatively little 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 less of an issue, since the agents do not realize their own reputations are being distorted. Still, I explore allowing for reputation concerns in the next subsection.

5.3. *Reciprocity with Partial Foresight.* If the agents had foresight, they would strategically consider a reputation effect in choosing current actions. In each t, R would maximize $x_{R,t} - \alpha_{R,t}(x_{R,t} - \tau_R)^2 + \pi(x_{R,t})$, with $\pi(x_{R,t})$ being the present value of future payoff effects of $x_{R,t}$.



Notes: Vertical axis $= E_{i,t}(s_{-i}) = s_{-i}^{i,t}$ versus time on horizontal axis. Time-frames of $t = 0 - 10$ (left graphs) and $t = 0 - 1,000$ (right graphs). One hundred simulations per set of variance parameter values; mean $s = \mu_s = 1$, mean tastes: $\mu_{\tau_R} = 1 = -\mu_{\tau_L}$, $b = 0.5$ in all graphs (L's initial prior mean for τ_R is $\tau_R^{L,0} = \mu_R - b = 0.5$ and R's initial prior mean for τ_L is $\tau_L^{R,0} = \mu_L + b = -0.5$. In lower graphs precision for τ's is relatively high.

Figure 2

simulated evolutions of i's expectations of the$-i$'s selfishness with false consensus bias for different time-frames and precision of beliefs about the opposition's tastes [color figure can be viewed at wileyonlinelibrary.com]

The $\pi()$ function can be complex depending on the degree of foresight. In order to relax the assumption of full myopia in a tractable way, I assume that each agent i, in each period, thinks just one period into the future in choosing the current action, and that there is common knowledge about this degree of foresight. The benefits of i's action in $t+1$ are orthogonal to those of her action in t, whereas the effects of the other agent's $t+1$ action are influenced by i's current action via its effect on $s_i^{-i,t+1}$. Thus, each agent now seeks to maximize

$$(18) \qquad \tilde{u}_{i,t}(x_{i,t}) = u_i^p(x_{i,t}) + \alpha_{i,t}u^s(x_{i,t};\tau_i) + \tilde{u}_i^p(x_{-i,t+1}(x_{i,t})) + \alpha_{i,t}\tilde{u}^s(x_{-i,t+1}(x_{i,t});\tau_i).$$

I write $\tilde{u}$ because the one-period-of-foresight assumption implies that i's perception of the next period's events is an approximation (i in t does not consider how $-i$ in $t+1$ considers how i in $t+2$ considers $-i$ in $t+3$, etc.).

Letting $x_{L,t+1}^{R,t-1}(x_{R,t})$ denote R's expectation of $x_{L,t+1}$ at the start of t (i.e., with information through $t-1$) given R's action $x_{R,t}$, R seeks to maximize

$$(19) \qquad \tilde{u}_{R,t}(x_{R,t}) = x_{R,t} - \alpha_{R,t}(x_{R,t} - \tau_R)^2 + x_{L,t+1}^{R,t-1}(x_{R,t}) - \alpha_{R,t}(x_{L,t+1}^{R,t-1}(x_{R,t}) - \tau_R)^2.$$

Let $\phi_t := \frac{\partial x_{L,t+1}^{R,t-1}(x_{R,t})}{\partial x_{R,t}}$, and maintain the $r = 1$ assumption (results are similar for $r < 1$). Then, the standard first-order condition implies:

$$(20) \qquad x_{R,t}^* = (1+\phi_t)\left(s_R + s_L^{R,t-1} + \tau_R\right) - \phi_t x_{L,t+1}^{R,t-1}(x_{R,t}^*).$$

Adding the reputation concern tempers $x_{R,t}$, since $\phi_t < 0$ (this is shown in the appendix, but should be intuitively clear since higher $x_{R,t}$ signals higher s_R to L, reducing $\alpha_{L,t+1}$, and thus making $x_{L,t+1}$ more consistent with L's private interests, i.e., more negative) and $x^{R,t-1}_{L,t+1}(x_{R,t})$ is typically negative.

However, this does not mean reputation concerns necessarily temper biased character inferences. Although there is correct common knowledge of ϕ_t due to common knowledge of the degree of foresight, the agents have different beliefs about $x^{R,t-1}_{L,t+1}(x_{R,t})$. In fact, L would expect this to be more extreme (more negative) than R because L believes R's belief about τ_L is more extreme than it really is. That is, L overestimates R's perceived marginal period $t+1$ social cost of $x_{R,t}$ on $x_{L,t+1}$ due to unawareness of R's false consensus bias. Thus, L expects too much foresight-induced temperance of R's action, implying the following.

Corollary 1. *The expected and strong affective polarization results stated in Proposition 3 continue to hold with the one-period-of-foresight assumption.*

The proof is in the Appendix; the intuition is the following. There are two future effects of $x_{R,t}$ on $\tilde{u}_{R,t}(x_{R,t})$: (1) the effect on R's private payoff; (2) the effect on R's weighted social payoff. L understands the first effect perfectly and thus accounts for it perfectly, in adjusting $\hat{x}_{R,t}$ to update beliefs about s_R. L does not understand the second effect perfectly; L underestimates R's expectation of s_L and overestimates R's expectation of L's expectation of s_L. But these errors perfectly cancel for reasons already discussed. L also overestimates R's belief about the extremism of τ_L. A more extreme τ_L increases the marginal cost to R from increasing the $s_R^{L,t+1}$ component of $x_{L,t+1}$ in $\tilde{u}^s(x_{L,t+1};\tau_R)$. This should push R to choose lower $x_{R,t}$, and causes L to infer that s_R is higher from a given observation of $\hat{x}_{R,t}$.

Put more simply, L thinks that R knows that L has a taste for left-leaning policies. R can temper this policy in the future with a more moderate policy now. If R does not realize how extreme L's tastes are, R will not temper R's current choice as much as L expects, and L will conclude that R must be more self-serving to justify this superficially relatively extreme and self-serving action. Consequently, the expression for the difference in first-order and second-order expectations driving the strong affective polarization result (b_t) is not decreased due to foresight, and so this result still holds. (The λ's may change due to foresight, but can still take any values in (0,1) for given σ's.)

This intuition implies that foresight could possibly even enhance affective polarization. This may be a stretch; regardless, this exercise does show that myopia alone certainly does not drive the results of Proposition 3. Even forward-looking agents can be misunderstood and have misunderstanding compound. Foresight alone does not necessarily mitigate this if the foresight is understood, and so each agent accounts for foresight's effects on the other agent's action appropriately. Foresight can even interact with other bias to exacerbate dislike.

6. JOINTLY CHOSEN ACTIONS AND LIMITED STRATEGIC THINKING

The preceding analysis and results raise two questions. First, what about other unmotivated biases—are the strong affective polarization results unique to the false consensus bias, or could there be other causes? Second, what about jointly, as opposed to unilaterally, determined actions?

In order to address these concerns, I now assume a single action, x, is chosen jointly in each period, with $x = x(x_L, x_R) = (x_L + x_R)/2$, and consider both strategically sophisticated behavior, and a new bias, limited strategic thinking. In order to simplify, and focus on strategic considerations, assume that $\tau_i = 0$ for each i and this is common knowledge. The myopic stage-game payoffs are then:

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

Since jointly chosen x endogenously creates a reciprocity motive (if player i prefers larger x, she has an incentive to increase x_i as she expects x_{-i} to decline and vice versa) it is natural to restrict r to now equal zero. In order to simplify algebra, drop the $(1/2)$ from the definition of α_i, so $\alpha_i = 1/s_i$ for each i, implying the best response functions are now:

$$\begin{aligned} x^*_{L,t}(x_{R,t}) &= -s_L - x_{R,t}, \\ x^*_{R,t}(x_{L,t}) &= s_R - x_{L,t}. \end{aligned} \tag{21}$$

There is no (Bayesian) Nash equilibrium (BNE) in the stage game, for any distribution of s_i or beliefs about the distribution of s_i, given that s_i has strictly positive support for each i. For any 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. For example, suppose there was common knowledge that the s_i's were both small, say 0.1. Suppose R believed $x_{L,t} = -1$. Then R's best reply would be $0.1 - (-1) = 1.1$. L's best reply to $x_{R,t} = 1.1$ is $-0.1 - (1.1) = -1.2$. R's best reply to this action is 1.3. The unraveling continues, with no fixed point.[17] The logic holds when there is uncertainty about s_{-i} so long as its expected value is strictly positive.

If x_L and x_R were each chosen from a closed, bounded, symmetric (around zero) interval, $[-a, a]$ for $a > 0$, then there would be a stage game BNE in which the players would choose opposite extreme actions: $x_L = -a$ and $x_R = a$, and neither player would update beliefs about the other player's s parameter. That is, although relatively extreme actions would be taken, they would occur in the first period and every period, and there would be no growth in dislike. This would be the case whether or not the players are subject to a bias, such as the false consensus bias.

Thus, in this model, simply making the actions chosen jointly and strategically cannot explain growth in extremism of actions or affective polarization. Next, consider behavioral strategic thinking, in particular, the level-k model of limited strategic thinking. This has become the benchmark behavioral alternative to equilibrium (Crawford et al., 2013). A level k 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 nonstrategic. 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. I use it because it is a tractable model for capturing a key issue—overconfidence in one's understanding of the other player's strategic behavior. Each level k player is certain in her understanding of this. In a two-player game, one player's beliefs about the other's strategic thinking can be correct, but it is impossible for both to be correct—and it is possible for both to be incorrect, and this is the case if the players "think" at the same level.[18]

Let $x^{\mathcal{L}k}_{i,t}$ denote the best response of player i when she is a level k thinker in period t. For $k > 0$, these are:

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

[17] 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.

[18] Regarding political interactions, although these are repeated, each one is unique and may involve different actors, and voters do not get direct feedback on beliefs they form on these interactions. It is not implausible that strategic motives could be consistently underappreciated by voters. Roughly speaking, this is what much of the public choice literature is about. Applying the level k model to voter beliefs is novel, to my knowledge, but very similar ideas are studied in the theory work on limited strategic thinking by voters of Szembrot (2017) and Demange and Van Der Straeten (2017), and supported experimentally by Szembrot (2018). Informally this idea is similar to the aphorism "don't hate the player, hate the game," which alludes to seemingly "bad" actions taken by individuals being explained by subtle strategic motives; see Washington (2013) for a piece on this topic titled with this phrase.

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, whereas this would typically not be the case for Level 1 thinkers. In the appendix, I show that results are similar for higher level strategic thinkers.

How do 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), or randomize with this mean. 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 $x_{L,t}^{\mathcal{L}2} = -s_L - s_R^{L,t-1}$, $x_{R,t}^{\mathcal{L}2} = s_R + s_L^{R,t-1}$. Note that this implies that assuming that level 2 players are myopic is without loss of generality: Since a level 2 player thinks her opponent $-i$ simply plays a function of s_{-i}, a level 2 player will not be concerned about the effect of her action on her opponent's beliefs.[19] A level-2 player L will therefore update her expectation of s_R as follows:

$$s_R^{L,t} = \lambda_t \mu_s + (1-\lambda_t)\left(\frac{1}{t}\right)\left(\hat{x}_1^R + \hat{x}_2^R + \cdots \hat{x}_t^R\right)$$

$$(22) \qquad = \lambda_t \mu_s + (1-\lambda_t)\left(\frac{1}{t}\right)\left(\left(s_R + s_L^{R,0}\right) + \left(s_R + s_L^{R,1}\right) + \cdots + \left(s_R + s_L^{R,t-1}\right) + \sum_{i=1}^{t} \epsilon_i^R\right).$$

R's actions are taken at face value as unbiased signals of s_R. $s_L^{R,1}, s_L^{R,2}$, etc, will be determined analogously as (increasing) functions of s_L and $s_R^{L,t}$'s for prior periods, which will in turn be functions of s_R and earlier $s_L^{R,t}$'s. Note that a fundamental difference between this case and the case of false consensus and out-group biases are that we can no longer use the difference in first-order and second-order expectations as a measure of bias, because second-order expectations are no longer unbiased.[20] A level 2 player i thinks that $-i$ is level 1 and therefore does not think i acts strategically and so i's actions are not used by $-i$ to update beliefs about s_i. In order 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 (22) implies

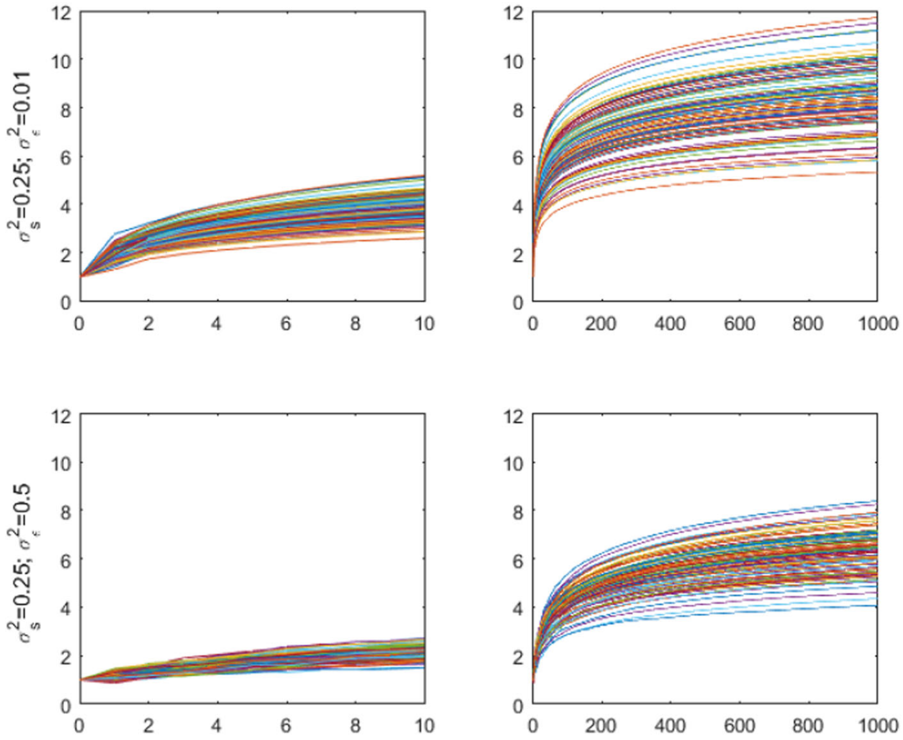
$$E_0\left(s_R^{L,t} | s_L, s_R\right) \geq \lambda_t \mu_s + s_t, \text{ in which}$$

$$(23) \qquad s_t = (1-\lambda_t)\left(s + \frac{1}{t}\sum_{i=1}^{t-1} s_i\right) \quad \text{for } t > 1 \text{ with } s_1 = (1-\lambda_1)s.$$

In the Appendix, this inequality is used to prove the following results.

[19] The myopia assumption also does not affect results for higher level players. A higher level L will think that the current action will make R infer that L is more selfish, making R's next action more extreme. But L thinks that by choosing an even more extreme action himself in the next period, she can compensate precisely for this increase R's extremism. So L thinks there is no future cost to current extremism, and so optimal actions are the same whether or not there is myopia.

[20] Moreover, as shown in the Appendix, level k players will in general have increasing second-order expectations; that is, level k players will think that they are becoming increasingly disliked by the opposition.



NOTES: Vertical axis = $E_{i,t}(s_{-i}) = s_{-i}^{i,t}$ versus time on horizontal axis. Time-frames are $t = 0 - 10$ (left graphs) and $t = 0 - 1000$ (right graphs). One hundred simulations per set of variance parameter values; prior mean for $s_i = \mu_s = 1$ in all graphs. Bottom graphs are for case of relatively high noise in observed actions (ϵ).

FIGURE 3

SIMULATED EVOLUTIONS OF i'S EXPECTATIONS OF $-i$'S SELFISHNESS WITH JOINT CHOICE OF x AND LEVEL 2 PLAYERS, FOR DIFFERENT TIME-FRAMES AND PRECISION OF OBSERVED ACTIONS [COLOR FIGURE CAN BE VIEWED AT WILEYONLINELIBRARY.COM]

PROPOSITION 4. *With common knowledge of tastes and $\tau_L = \tau_R = 0$, $r = 0$, and jointly chosen x, if L and R are $\mathcal{L}2$ thinkers who assume $\mathcal{L}0$ thinkers play $x_i = 0$, 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 3 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 3 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 response to $x_R = 0$), but L plays $s_R^{L,0} + s_L = \mu_s + s_L$ (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 ($\sigma_s^2 \geq \sigma_\epsilon^2$). 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.

The similarity of the false consensus and level k results because the underlying mechanism is perhaps surprisingly similar: In both cases, each player does not observe two parameters of the other player (s_{-i} and either tastes or level of strategic thinking). Beliefs about one of the parameters, s_{-i}, become biased upward due to biased beliefs about the other parameter. This leads to a more extreme action, which is interpreted as evidence that s_{-i} is even higher due to unawareness of the initial bias, and a snowball effect ensues. This mechanism could apply with other specific biases, as noted at the end of Section 1.

7. CONCLUDING REMARKS

Why does disagreement lead to dislike? Why do escalation of extremism and hostility often go hand-in-hand, in party politics and beyond? In this article, I present a model that shows how these phenomena can be caused by a combination of Bayesian inference and biases in priors that seem unrelated to character.

To be clear, I certainly do not claim that the factors studied in this article are the primary cause of partisan affective polarization. The article also does not imply that individuals on both sides are equally responsible for this phenomenon or have equally valid views.[21] My suggested interpretation of this article is that the biases studied here, and perhaps similar biases, are neglected contributing factors to hard feelings in politics, and other repeated bilateral settings that have gone downhill.

A question that the analysis raises, but fails to address, is what has caused affective polarization to be particularly intense in the United States in recent years. Partisan hostility in the United States may have been relatively low before this period due to the depolarizing effects of having a "common enemy" during World War II and the Cold War. Growth in hostility may have been especially intense since then due to the low baseline, and at least two other major factors. First, there has been a confluence of institutional changes leading to more partisan behavior by U.S. politicians, such as increased gerrymandering, partisan sorting (liberal Republicans becoming Democrats, and conservative Democrats switching to Republicans), and changes in norms and rules in Congress leading to increased use of the filibuster and other relatively aggressive tactics (Barber and McCarty, 2015). As voters have observed politicians from the other party taking seemingly more extreme actions for these exogenous reasons, voters might have made more negative character inferences, strengthening the hostility-extremism cycle.[22] Second, the changing media environment has likely interacted with partisan affect and behavior. Cable news has experienced especially significant changes in the United States in the last few decades, with the introduction of Fox News in the mid-90s and repositioning of MSNBC in the 2000s (Martin and Yurukoglu, 2017). And, although online media has developed around the world in recent years, online media effects may have been stronger in the United States due to interactions with cable news and institutional political changes. It is also worth noting that partisan divides are now more severe than those of religious and ethnic groups across a range of nations (Westwood et al., 2015), suggesting that affective polarization has also occurred in other nations.

Another question that warrants discussion is whether it is plausible that such extreme bias about character could occur and persist. I am confident here asserting that the answer is yes: Although errors in beliefs tend to be eroded in situations in which people get frequent feedback on such errors, this is not the case for many types of beliefs. Large fractions of populations can hold wildly inaccurate beliefs, and can continue to hold such beliefs indefinitely if there

[21] See Greene (2014) for an important discussion of asymmetry in the validity of partisan perspectives, including a response to Haidt (2012), and Ditto et al. (2018) for a meta-study that concludes that biased interpretation of new information is similar across the political spectrum.

[22] For example, if some liberal Republicans switched party affiliations in the 1980s and 1990s and became Democrats, causing the remaining Republicans to be less moderate on average and thus less likely to compromise on legislation simply due to ideological principles, Democrats may have misinterpreted this less cooperative behavior as a sign of bad character. This may have led to retaliation, leading to retaliation in turn from Republicans, etc.

is no direct cost to doing so,[23] and especially if there are social or psychological benefits to maintaining such beliefs (Akerlof, 1989; Caplan, 2011). The fact that hostility may be Bayesian and based on a great deal of information may also contribute to why it is so hard to dispel in reality.

An aspect of the false consensus results that is most questionable is that the players are unaware of how much they are disliked by the opposition. Perhaps individuals ignore or are oblivious to the hostility felt toward them by the opposition, or perceive this hostility as exaggerated for strategic or psychological reasons. An alternative explanation of this aspect of the results is that people are aware of being disliked by those on the other side, but attribute it to being 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, and this is what makes R resent and dislike L. This alternative mechanism is not mutually exclusive with that of the model and so would not invalidate the model. It is also worth noting again that this result does not hold for level k players—they are, in general, aware of becoming increasingly disliked over time.

Are there any policies that could mitigate biased affective polarization? I do not claim to have any sure-fire answers here, but still offer a few thoughts. One possibility is that simply spreading awareness of this research, and similar research, could be effective. As people learn that partisan hostility is often driven by cognitive bias, this may stigmatize hostility and reduce its prevalence. A related possibility would be to develop a norm of respect for the opposition's motives, perhaps led by political figures with relatively high levels of bipartisan support.[24] Another, more costly, option would be to introduce public education programs. Perhaps public schools could include education on political psychology and cognitive bias, or just on empathy more broadly (Gehlbach, 2017), as part of civics programs. Policies that promote fact-checking and diversity of viewpoints in political media could also be helpful. A final, particularly impractical but perhaps intriguing, option would be to build off the analogy to troubled marriages, and find arbiters from relatively neutral nations to mediate partisan disputes, akin to marriage counselors. Regardless, improving understanding of the issue is important in its own right, and could enhance the ability to address the problem in unforeseen ways.

APPENDIX: PROOFS

A.1. Proof of Proposition 1.

PROOF. In order 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 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 3.

PROOF. In order to prove the first part, note expected polarization occurs because

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

[23] For example, a recent survey found that 45% of Americans believe in ghosts (https://www.usatoday.com/story/news/nation-now/2017/10/25/how-many-people-believe-ghosts-dead-spirits/794215001/).

[24] Joe Biden discusses related ideas in the March 28, 2018, episode of "Pod Save America."

$$= \mu_s + E_0\left((1-\lambda_t)\left(\frac{1}{t}\sum_{i=1}^{t-1} b_i + b\right)\right),$$

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

In order 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* $x_t = H_t = 1 + 1/2 + 1/3 + \cdots + 1/t$ *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 $x_{t+1} = H_{t+1}$.

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

■

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 $b > 0$.

In order to prove the third part, note

$$\begin{aligned}
&E_0(s_R^{L,t} - s_R^{L,t-1} | s_L, s_R) = \\
&\Delta\lambda_t(\mu_s - s_R - b) + \left((1-\lambda_t)\left(\frac{1}{t}\sum_{i=1}^{t-1} b_i\right) - (1-\lambda_{t-1})\left(\frac{1}{t-1}\sum_{i=1}^{t-2} b_i\right)\right) = \\
&\Delta\lambda_t(\mu_s - s_R - b) + (1-\lambda_t)\left(\frac{1}{t}\sum_{i=1}^{t-1} b_i - \frac{1}{t-1}\sum_{i=1}^{t-2} b_i\right) \\
&+ (1-\lambda_t)\left(\frac{1}{t-1}\sum_{i=1}^{t-2} b_i\right) - (1-\lambda_{t-1})\left(\frac{1}{t-1}\sum_{i=1}^{t-2} b_i\right) = \\
&\Delta\lambda_t\left(\mu_s - s_R - b - \frac{1}{t-1}\sum_{i=1}^{t-2} b_i\right) + (1-\lambda_t)\left(\frac{1}{t}\sum_{i=1}^{t-1} b_i - \frac{1}{t-1}\sum_{i=1}^{t-2} b_i\right). \qquad \text{(A.2)}
\end{aligned}$$

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

A.3. Robustness of Proposition 1 to Uncertainty in τ_{-i}. Maintain the out-group bias assumption of Section 4: $E_{i,0}(s_{-i}) = \mu_s + b$ with $b > 0$ and unawareness of bias, so $E_{i,-i,0}(s_i) = \mu_s$, and assume that each agent i does not observe τ_{-i} but has unbiased priors about this parameter. Then, by the logic discussed in Section 5, at the end of $t = 1$, L will update via:

$$\begin{aligned} s_R^{L,1} &= \lambda_1 s_R^{L,0} + (1-\lambda_1)(\hat{x}_{R,1} - r s_L^{L,R,0} - \tau_R^{L,0}) \\ &= \lambda_1(\mu_s + b) + (1-\lambda_1)(\hat{x}_{R,1} - r\mu_s - \mu_{\tau_R}), \end{aligned} \tag{A.3}$$

and R believes that L updates with

$$\begin{aligned} s_R^{R,L,1} &= \lambda_1 s_R^{R,L,0} + (1-\lambda_1)(\hat{x}_{R,1} - r s_L^{R,L,R,0} - \tau_R^{R,L,0}) \\ &= \lambda_1\mu_s + (1-\lambda_1)(\hat{x}_{R,1} - r(\mu_s + b) - \mu_{\tau_R}). \end{aligned} \tag{A.4}$$

Consequently, just as in Section 4, the difference in expectations is

$$s_R^{L,1} - s_R^{R,L,1} = \lambda_1 b + (1-\lambda_1) r b = b_1, \tag{A.5}$$

with b_t as defined in Subsection 5.2, and note now $b_0 = b$. That is, the uncertainty in τ_{-i} does not affect the difference in first-order and second-order expectations at the end of period 1. In order to determine b_t for $t > 1$, we can use the logic discussed in Subsection 5.2 (in particular, the first line of equation (15), using the r parameter rather than $r = 1$), to see that[25]

$$b_t = \lambda_t b + (1-\lambda_t)\left(\frac{r}{t}\sum_{i=0}^{t-1} b_i\right) \text{ for } t \geq 1. \tag{A.6}$$

It is straightforward to see that if $r < 1$, then since $b_1 < b_0$ and $1 - \lambda_t$ increases in t, by induction b_t is decreasing in t. If $r = 1$, then $b_t = b$ for all t, just as in Section 4.

In contrast to that section, however, λ_t no longer converges to zero, and so second-order expectations do not converge to truth but instead to a convex combination of the prior and the true value plus the difference between τ_{-i} and $E_0(\tau_{-i})$, as referred to in parts 3 and 2 of Propositions 2 and 3, respectively, due to the identification problem (the inability to distinguish between selfishness and tastes even in the long run): $\frac{\sigma_\tau^2}{\sigma_s^2+\sigma_\tau^2}\mu_s + \frac{\sigma_s^2}{\sigma_s^2+\sigma_\tau^2}(s_R + \tau_R - \mu_{\tau_R})$. But again, if $r = 1$, then first-order expectations are greater than second-order by exactly b for all t, and, if $r < 1$, by an amount less than b for all $t > 0$ that decreases over t, and so in both cases neither type of affective polarization occurs.

In order to see the intuition, note that if τ_R is very large, and L does not observe this directly and only observes R's (large) actions, then L's posterior expectation of s_R will in general be too high, and thus L's actions will be too extreme due to reciprocity. But R perfectly understands this, and therefore L's reciprocity will not affect R's beliefs about s_L, and R's action will not become more extreme in retaliation. Consequently, uncertainty about τ unto itself does not lead to any exacerbation of biased beliefs about s_{-i} over time. Put differently, it is not the first-order uncertainty that causes bias to build upon itself, but the difference in first-order and second-order beliefs. This difference occurs when one player has a bias that the other does not fully appreciate, and the difference does not occur merely due to first-order uncertainty.

A.4. Proof of Corollary 1.

PROOF. Expressing L's updating based on the mean of observed R actions is not helpful for simplifying the analysis in this case given that the expectations of effects on next period's

[25] The first line of equation (15), with the $\frac{1}{2}$ multiplied by r, minus the analogous version for the second-order expectations, yields the derived expression here.

actions would have to be accounted for in each period, and these expectations are functions of updated means for the τ's, as will be seen shortly. Thus, consider the updating of s_R from period to period, using updated prior means of the τ's. Rearranging (20) to obtain the adjusted signal that L uses for updating beliefs about s_R, yields:

$$\begin{aligned} s_R^{L,t} &= \lambda_t s_R^{L,t-1} + (1-\lambda_t)\,(\hat{x}_{R,t}/(1+\phi_t\,) \\ &\quad + \,(\phi_t/(1+\phi_t))x_{L,t+1}^{L,R,t-1}(\hat{x}_{R,t}) - s_L^{L,R,t-1} - \tau_R^{L,t-1}\Big)\,. \end{aligned} \tag{A.7}$$

Here, $x_{L,t+1}^{L,R,t-1}(\hat{x}_{R,t})$ refers to L's expectation of $x_{L,t+1}^{R,t-1}(x_{R,t})$; the $\hat{x}_{R,t}$ and $x_{R,t}$ are essentially interchangeable due to linearity. Given the one-period-of-foresight assumption, $x_{L,t+1}^{R,t-1}(x_{R,t}) = -s_L^{R,t-1} - s_R^{R,L,t} + \tau_L^{R,t-1}$ (this is myopic L's optimal action in $t+1$) with $s_R^{R,L,t}$ being a function of $x_{R,t}$. Specifically, $x_{L,t+1}^{R,t-1}(x_{R,t}) =$

$$-s_L^{R,t-1} - \lambda_t s_R^{R,L,t-1} - (1-\lambda_t)\left(\frac{1}{1+\phi_t}\left(x_{R,t} + \phi_t x_{L,t+1}^{R,L,R,t-1}(x_{R,t})\right) - s_L^{R,L,R,t-1} - \tau_R^{R,L,t-1}\right) + \tau_L^{R,t-1}.$$

Note that $\hat{x}_{R,t}$ is replaced by $x_{R,t}$ because R does not observe $\hat{x}_{R,t}$ prior to t and R's expectation of $\hat{x}_{R,t}$ is $x_{R,t}$. Unawareness of bias implies third-order beliefs equal first-order beliefs, so $x_{L,t+1}^{R,L,R,t-1}(x_{R,t}) = x_{L,t+1}^{R,t-1}(x_{R,t})$, etc. This equation is thus linear in $x_{L,t+1}^{R,t-1}(x_{R,t})$, and can be solved for

$$\begin{aligned} &x_{L,t+1}^{R,t-1}(x_{R,t}) \\ &= \psi_t\Big(-\lambda_t s_L^{R,t-1} - \lambda_t s_R^{R,L,t-1} - (1-\lambda_t)x_{R,t}/(1+\phi_t) + (1-\lambda_t)\tau_R^{R,L,t-1} + \tau_L^{R,t-1}\Big), \end{aligned} \tag{A.8}$$

with $\psi_t = 1/(1+(1-\lambda_t)(\phi_t/(1+\phi_t)))$, which the agents have correct common knowledge about. Since $\phi_t = \frac{\partial x_{L,t+1}^{R,t-1}(x_{R,t})}{\partial x_{R,t}} = -\psi_t(1-\lambda_t)/(1+\phi_t)$, it can be shown that if $\phi_t > 0$, this would imply a contradiction (since ψ_t would also have to be > 0, implying $\phi_t < 0$) and if $\phi_t < -1$ this would also imply a contradiction (this would also imply $\psi_t > 0$ implying $\phi_t > 0$). Thus, $\phi_t \in (-1, 0)$ and ψ_t must be > 0 (if not, this would again imply a contradiction).

Again, second-order expectations are equal to unbiased expectations, and so we can look at the difference between first- and second-order expectations to obtain a measure of bias for the former: $s_R^{L,t} - s_R^{R,L,t} = \lambda_t(s_R^{L,t-1} - s_R^{R,L,t-1}) + (1-\lambda_t)((\phi_t/(1+\phi_t))(x_{L,t+1}^{L,R,t-1}(x_{R,t}) - x_{L,t+1}^{R,t-1}(x_{R,t})) - (s_L^{L,R,t-1} - s_L^{R,t-1}) - (\tau_R^{L,t-1} - \tau_R^{R,L,t-1}))$. (The first $\hat{x}_{R,t}$ terms from the right-hand side of (A.7), a form of which is used in both first-order and second-order expectations, cancel right away; the second $\hat{x}_{R,t}$ terms (the components of the $x_{L,t+1}^{R,t-1}(x_{R,t})$ terms) will be discussed soon.) Again, this expression is the same as for the myopic case except for the single term involving ϕ. If the sign of this term is consistent with the sign of the other effects, then this new term simply reinforces the other effects, and this would complete the proof. (Bias against the other agent would always be higher after the first period (as compared to the myopic case), and subsequent bias increases as a function of initial bias in the same way for both the cases of myopia and foresight.) $(\phi_t/(1+\phi_t))$ is multiplied by $(x_{L,t+1}^{L,R,t-1}(x_{R,t}) - x_{L,t+1}^{R,t-1}(x_{R,t})) = \psi_t((1-\lambda_t)(\tau_R^{L,t-1} - \tau_R^{R,L,t-1}) + (\tau_L^{L,R,t-1} - \tau_L^{R,t-1}))$. The right-hand side simplifies to this expression because the $x_{R,t}$ terms from (A.8) become $\hat{x}_{R,t}$ and are the same and thus cancel, and the difference in the $-\lambda_t s_L^{R,t-1} - \lambda_t s_R^{R,L,t-1}$ terms also completely cancel because of symmetry in any bias (and $r = 1$). Thus, any bias is driven by the differences in expectations of the τ's. Both of these differences are unambiguously negative. And, they are both multiplied by $(\phi_t/(1+\phi_t)) < 0$. Thus, the new effect is positive, consistent with the other effects. Thus, $s_R^{L,t} - s_R^{R,L,t}$ can only be increased by the forward-looking behavior, as compared to the case of full myopia. ■

A simplified way to see the intuition, and mechanics of the analysis, is as follows. Suppose $x_{L,t+1}^{R,t-1}(x_{R,t}) = -(\gamma_t + \delta_t x_{R,t}) + \tau_L$ with $\gamma_t > 0$ and $\delta_t \in (0, 1)$ for all t, and so $\phi_t = -\delta_t \in (-1, 0)$. That is, R believes that L takes an action with a constant (for the time period) negative component, $-\gamma_t$, and a component that is negatively linear in $x_{R,t}$, $-\delta_t x_{R,t}$, plus L's tastes parameter (which is also likely negative) as in the standard myopic case. Suppose also that this functional form, and the γ and δ parameters, are common knowledge. Then,

$$x_{R,t}^* = (1+\phi_t)\left(s_R + s_L^{R,t-1} + \tau_R\right) - \phi_t\left(-(\gamma_t + \delta_t x_{R,t}^*) + \tau_L^{R,t-1}\right)$$

$$x_{R,t}^* = (1/(1-\phi_t\delta_t))\Big((1+\phi_t)\left(s_R + s_L^{R,t-1} + \tau_R\right) - \phi_t\left(-\gamma_t + \tau_L^{R,t-1}\right)\Big). \tag{A.9}$$

Thus,

$$\hat{x}_{R,t} = (1/(1-\phi_t\delta_t))\Big((1+\phi_t)(s_R + s_L^{R,t-1} + \tau_R) - \phi_t\left(-\gamma_t + \tau_L^{R,t-1}\right)\Big) + \epsilon_{R,t} \tag{A.10}$$

and L updates with

$$\begin{aligned} s_R^{L,t} &= \lambda_t s_R^{L,t-1} \\ &\quad + (1-\lambda_t)\left(((1-\phi_t\delta_t)/(1+\phi_t))\hat{x}_{R,t} - \delta_t\phi_t/(1+\phi_t)\right. \\ &\quad + \left.\phi_t/(1+\phi_t)\tau_L^{L,R,t-1} - s_L^{L,R,t-1} - \tau_R^{L,t-1}\right). \end{aligned} \tag{A.11}$$

Both players hold the same beliefs about the two terms involving δ_t in each period, so they cancel in $s_R^{L,t} - s_R^{R,L,t}$. This represents the "neutrality" of whether or not players have foresight—if they share an understanding of how this foresight affects current actions, then it does not affect character inferences. However, the difference, $(1-\lambda_t)\phi_t/(1+\phi_t)(\tau_L^{L,R,t-1} - \tau_L^{R,t-1})$, is nonneutral, and is actually strictly positive, since both $\phi_t/(1+\phi_t)$ and $(\tau_L^{L,R,t-1} - \tau_L^{R,t-1})$ are strictly negative. Because R underestimates the extremism of L's tastes, and L is not aware of this, R will not temper her current action (to temper L's future action) as much as L expects R to. This will make L further overestimate s_R based on R's current action. And, the last two terms in $s_R^{L,1}$ (and $s_R^{L,1} - s_R^{R,L,1}$) are the same as those of the myopic case. Thus, $s_R^{L,1} - s_R^{R,L,1}$ will be driven upward, as compared to the myopic case (without loss of generality ignoring the differences in λ's between the myopic/foresight cases). This difference will in fact be $(1-\lambda_1)b(\delta_t/(1-\delta_t)+1)$ (since $-\phi_t/(1+\phi_t) = \delta_t/(1-\delta_t)$), which is strictly greater than $(1-\lambda_1)b$ for the myopic case. Since later differences in these expectations $(s_R^{L,t} - s_R^{R,L,t})$ are increasing functions of earlier differences, which is what drives the strong affective polarization result, this result will also hold for the model with foresight.

A.5. *Proof of Proposition 4.*

PROOF. Expected polarization occurs because

$$E_0(s_{-i}^{i,t}) = \mu_s + m_t, \text{ in which}$$

$$m_t = (1-\lambda_t)\left(\mu_s + \frac{1}{t}\sum_{i=1}^{t-1} m_i\right) \text{ for } t > 1 \text{ with } m_1 = (1-\lambda_1)\mu_s, \tag{A.12}$$

and m_t is increasing.

In order 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 $x_i = (t/(t+1))(1 + (\frac{1}{t})\sum_{i=1}^{t-1} x_i)$ for $t > 1$ with $x_1 = 1/2$. Then $x_t + 1 = H_{t+1} = 1 + 1/2 + 1/3 + \cdots 1/(t+1)$, 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

$$\begin{aligned}
x_i &= (t/(t+1))\left(1 + \frac{1}{t}(H_2 - 1 + H_3 - 1 + \cdots + H_t - 1)\right)\\
&= t/(t+1) + (1/(t+1))((1/2) + (1/2 + 1/3) + \cdots (1/2 + 1/3 + \cdots + 1/t))\\
&= t/(t+1) + (t-1)/(2(t+1)) + (t-2)/(3(t+1)) + \cdots + 1/(t(t+1))\\
&= t/(t+1) + (t+1-2)/(2(t+1)) + (t+1-3)/(3(t+1)) + \cdots\\
&\quad + (t+1-t)/(t(t+1))\\
&= t/(t+1) + 1/2 - 1/(t+1) + 1/3 - 1/(t+1) + \cdots 1/t - 1/(t+1)\\
&= t/(t+1) + (H_t - 1) - (t-1)/(t+1) = (H_t - 1) + 1/(t+1)\\
&= H_{t+1} - 1. \qquad \text{(A.13)}
\end{aligned}$$

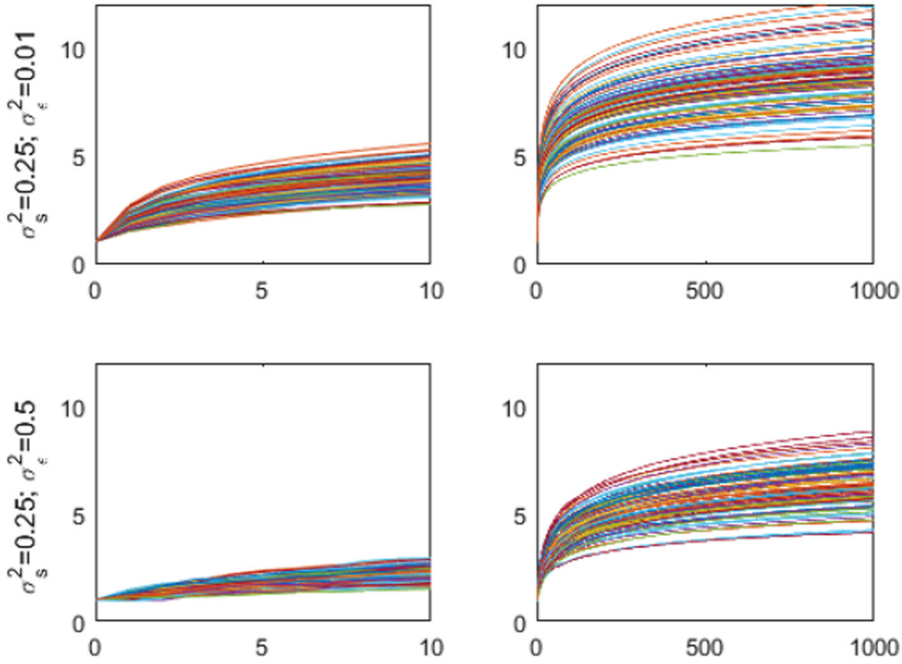
■

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 3. ■

A.6. Level 3 Strategic Thinking. It follows from the discussion in the text that $x_{L,t}^{\mathcal{L}3} = -s_L - s_R^{L,t-1} - s_L^{L,R,t-1}$, and $x_{R,t}^{\mathcal{L}3} = s_R + s_L^{R,t-1} + s_R^{R,L,t-1}$. Moreover, since level 3 players believe their opponents are level 2, a level 3 player L will update beliefs about s_R as follows:

$$\begin{aligned}
s_R^{L,t} &= \lambda_t\mu_s + (1-\lambda_t)\left(\frac{1}{t}\right)\left(\left(\hat{x}_1^R - s_L^{L,R,0}\right) + \cdots + \left(\hat{x}_t^R - s_L^{L,R,t-1}\right)\right)\\
&= \lambda_t\mu_s + (1-\lambda_t)\left(s_R + \frac{1}{t}\left(\left(s_L^{R,0} + s_R^{R,L,0} - s_L^{L,R,0}\right) + \cdots\right.\right.\\
&\quad \left.\left. + \left(s_L^{R,t-1} + s_R^{R,L,t-1} - s_L^{L,R,t-1}\right) + \sum_{i=1}^{t} \epsilon_i^R\right)\right). \qquad \text{(A.14)}
\end{aligned}$$

For each observed action by R, L subtracts off one expectation term (L's second-order expectation). But the action is the sum of s_R and two of R's expectations (R's first-order and second-order expectations), so again L is not accounting for R's strategic motives sufficiently. It appears that this would lead to a similar level of bias in interpreting R's actions as compared to the level 2 case, since again these actions are one expectation term greater than L expects. Moreover, even if the term that L subtracts off from R's action is large, since the subtracted off term is L's second-order expectation, if this term is larger, this would drive L's action to be more extreme. This would cause R's interpretation of L's action to be more biased (since this is the component of L's action that R fails to account for), driving R's first-order expectation and action upward, and creating another cycle of affective polarization. I do not work out the analytics of this case but simulate it and present results in Figure A1. Results are indeed very



NOTES: Vertical axis $= E_{i,t}(s_{-i}) = s^{i,t}_{-i}$ versus time on horizontal axis. Time-frames are $t = 0 - 10$ (left graphs) and $t = 0 - 1,000$ (right graphs). One hundred simulations per set of variance parameter values; prior mean for $s_i = \mu_s = 1$ in all graphs. Bottom graphs are for case of relatively high noise in observed actions (ϵ).

FIGURE A1

SIMULATED EVOLUTIONS OF i'S EXPECTATIONS OF $-i$'S SELFISHNESS WITH JOINT CHOICE OF x AND LEVEL 3 PLAYERS, FOR DIFFERENT TIME-FRAMES AND PRECISION OF OBSERVED ACTIONS [COLOR FIGURE CAN BE VIEWED AT WILEYONLINELIBRARY.COM]

similar to those of the level 2 case shown in Figure 3. The logic of analysis for higher level players is similar and so results should be similar as well.

REFERENCES

ACEMOGLU, D., V. CHERNOZHUKOV, AND M. YILDIZ, "Fragility of Asymptotic Agreement under Bayesian Learning," *Theoretical Economics* 11 (2016), 187–225.

———, AND A. WOLITZKY, "Cycles of Conflict: An Economic Model," *The American Economic Review* 104 (2014), 1350–67.

AKERLOF, G. A., "The Economics of Illusion," *Economics & Politics* 1 (1989), 1–15.

ANDREONI, J., AND T. MYLOVANOV, "Diverging Opinions," *American Economic Journal: Microeconomics* 4 (2012), 209–232.

AUMANN, R. J., "Agreeing to Disagree," *The Annals of Statistics* 4 (1976), 1236–1239.

BABCOCK, L., AND G. LOEWENSTEIN, "Explaining Bargaining Impasse: The Role of Self-Serving Biases," *The Journal of Economic Perspectives* 11 (1997), 109–126.

BALIGA, S., E. HANANY, AND P. KLIBANOFF, "Polarization and Ambiguity," *The American Economic Review* 103 (2013), 3071–83.

BARBER, M., AND N. MCCARTY, "Causes and Consequences of Polarization," *Political Negotiation: A Handbook* 37 (2015), 39–43.

BATSON, C. D., AND A. A. POWELL, "Altruism and Prosocial Behavior," *Handbook of Psychology* (2003), 463–484.

BELCO, M., AND B. ROTTINGHAUS, *The Dual Executive: Unilateral Orders in a Separated and Shared Power System* (Stanford, CA: Stanford University Press, 2017).

BENJAMIN, D. J., "Errors in Probabilistic Reasoning and Judgment Biases," Technical Report, National Bureau of Economic Research, 2018.

BENOÎT, J.-P., AND J. DUBRA, "When Do Populations Polarize? An Explanation," *International Economic Review* (2019). https://doi.org/10.1111/iere.12400.

Blomberg, B., and J. E. Harrington, “A Theory of Rigid Extremists and Flexible Moderates with an Application to the US Congress,” *The American Economic Review* 90 (2000), 605–20.

Bordalo, P., K. Coffman, N. Gennaioli, and A. Shleifer, “Stereotypes,” *The Quarterly Journal of Economics* 131 (2016), 1753–94.

Bradbury, T. N., S. R. Beach, F. D. Fincham, and G. M. Nelson, “Attributions and Behavior in Functional and Dysfunctional Marriages,” *Journal of Consulting and Clinical Psychology* 64 (1996), 569–576.

Bullock, J. G., “Partisan Bias and the Bayesian Ideal in the Study of Public Opinion,” *The Journal of Politics* 71 (2009), 1109–1124.

Butler, J. V., P. Giuliano, and L. Guiso, “Trust, Values, and False Consensus,” *International Economic Review* 56 (2015), 889–915.

Caplan, B., *The Myth of the Rational Voter: Why Democracies Choose Bad Policies* (Princeton, NJ: Princeton University Press, 2011).

Crawford, V. P., M. A. Costa-Gomes, and N. Iriberri, “Structural Models of Nonequilibrium Strategic Thinking: Theory, Evidence, and Applications,” *Journal of Economic Literature* 51 (2013), 5–62.

Curry, O. S., M. J. Chesters, and C. J. Van Lissa, “Mapping Morality with a Compass: Testing the Theory of ‘Morality as Cooperation’ with a New Questionnaire,” *Journal of Research in Personality* 78 (2018), 106–124.

Demange, G., and K. Van Der Straeten, “Communicating on Electoral Platforms,” *Journal of Economic Behavior & Organization* (2017). https://doi.org/10.1016/j.jebo.2017.03.006.

Ditto, P. H., B. S. Liu, C. J. Clark, S. P. Wojcik, E. E. Chen, R. H. Grady, J. B. Celniker, and J. F. Zinger, “At Least Bias Is Bipartisan: A Meta-Analytic Comparison of Partisan Bias in Liberals and Conservatives,” *Perspectives on Psychological Science* (2018), 17456916–17746796.

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

Fudenberg, D., G. Romanyuk, and P. Strack, “Active Learning With a Misspecified Prior,” *Theoretical Economics* 12.3 (2017), 1155–1189.

Garrett, K. N., and A. Bankert, “The Moral Roots of Partisan Division: How Moral Conviction Heightens Affective Polarization,” *British Journal of Political Science* (2018), 1–20.

Gehlbach, H., “Learning to Walk in Anothers Shoes,” *Phi Delta Kappan* 98 (2017), 8–12.

Glaeser, E. L., “The Political Economy of Hatred,” *The Quarterly Journal of Economics* 120 (2005), 45–86.

Greene, J., *Moral Tribes: Emotion, Reason, and the Gap Between Us and Them* (New York, NY: Penguin, 2014).

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

Heidhues, P., B. Kőszegi, and P. Strack, “Unrealistic Expectations and Misguided Learning,” *Econometrica* 86 (2018), 1159–1214.

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

Howell, W. G., *Power Without Persuasion: The Politics of Direct Presidential Action* (Princeton, NJ: Princeton University Press, 2003).

Iyengar, S., G. Sood, and Y. Lelkes, “Affect, Not Ideology a Social Identity Perspective on Polarization,” *Public Opinion Quarterly* 76 (2012), 405–31.

Jéheil, P., “Limited Horizon Forecast in Repeated Alternate Games,” *Journal of Economic Theory* 67 (1995), 497–519.

Klumpp, T., and H. M. Mialon, “On Hatred,” *American Law and Economics Review* (2013), aht004.

Lelkes, Y., “Mass Polarization: Manifestations and Measurements,” *Public Opinion Quarterly* 80 (2016), 392–410.

———, G. Sood, and S. Iyengar, “The Hostile Audience: The Effect of Access to Broadband Internet on Partisan Affect,” *American Journal of Political Science* 61 (2015), 5–20.

Levine, D. K., “Modeling Altruism and Spitefulness in Experiments,” *Review of Economic Dynamics* 1 (1998), 593–622.

Lewis, M., *The Undoing Project: A Friendship that Changed Our Minds* (New York, NY: Norton, 2016).

Loh, I., and G. Phelan, “Dimensionality and disagreement: Asymptotic Belief Divergence in Response to Common Information.” *International Economic Review* (2019). https://doi.org/10.1111/iere.12406.

Lowande, K. S., “The Contemporary Presidency After the Orders: Presidential Memoranda and Unilateral Action”, *Presidential Studies Quarterly* 44 (2014), 724–41.

Martin, G. J., and A. Yurukoglu, “Bias in Cable News: Persuasion and Polarization,” *American Economic Review* 107 (2017), 2565–99.

Mason, L., *Uncivil Agreement: How Politics Became Our Identity* (Chicago, IL: University of Chicago Press, 2018).

McMurray, J., “Ideology as Opinion: A Spatial Model of Common-Value Elections,” *American Economic Journal: Microeconomics* 9 (2017), 108–40.

Mercer, J., “Emotional Beliefs,” *International Organization* 64 (2010), 1–31.

Moshagen, M., B. E. Hilbig, and I. Zettler, “The Dark Core of Personality,” *Psychological Review* 125 (2018), 656–88.

Nussbaum, M. C., *Upheavals of Thought: The Intelligence of Emotions* (Cambridge, MA: Cambridge University Press, 2003).

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

Pessoa, L., “On the Relationship Between Emotion and Cognition,” *Nature Reviews Neuroscience* 9 (2008), 148–58.

Piketty, T., “Social Mobility and Redistributive Politics,” *The Quarterly Journal of Economics* 110 (1995), 551–84.

Pinker, S., *The Sense of Style: The Thinking Person’s Guide to Writing in the 21st Century!* (New York, NY: Penguin Books, 2015).

Ross, L., D. Greene, and P. House, “The False Consensus Effect: An Egocentric Bias in Social Perception and Attribution Processes,” *Journal of Experimental Social Psychology* 13 (1977), 279–301.

———, F. Sicoly et al.“Egocentric Biases in Availability and Attribution,” *Journal of Personality and Social Psychology* 37 (1979), 322–36.

Rudalevige, A., “As a Candidate, Trump Criticized Obama’s Use of Executive Power. So Guess What Powers President Trump Has Been Leaning On?” *The Washington Post* (1 2018).

Ruffle, B. J., and R. Sosis, “Cooperation and the In-Group–Out-Group Bias: A Field Test on Israeli Kibbutz Members and City Residents,” *Journal of Economic Behavior & Organization* 60 (2006), 147–63.

Ryan, T. J., “Reconsidering Moral Issues in Politics,” *The Journal of Politics* 76 (2014), 380–97.

Sethi, R., and M. Yildiz, “Public Disagreement,” *American Economic Journal: Microeconomics* 4 (2012), 57–95.

Snowe, O. J., “The Effect of Modern Partisanship on Legislative Effectiveness in the 112th Congress,” *Harvard Journal on Legislation* 50 (2013), 21.

Stone, D. F., “Unmotivated Bias and Partisan Hostility: Empirical Evidence,” *Journal of Behavioral and Experimental Economics* 79 (2019), 12–26.

Szembrot, N., “Are Voters Cursed When Politicians Conceal Policy Preferences?” *Public Choice* 173 (2017), 25–41.

———, “Experimental Study of Cursed Equilibrium in a Signaling Game,” *Experimental Economics* 21.2 (2018), 257–291.

Tajfel, H., “Social Psychology of Intergroup Relations,” *Annual Review of Psychology* 33 (1982), 1–39.

Uhlmann, E. L., D. A. Pizarro, and D. Diermeier, “A Person-Centered Approach to Moral Judgment,” *Perspectives on Psychological Science* 10 (2015), 72–81.

Washington, R., “Don’t Hate the Player, Hate the Game: Remembering James Buchanan,” *The Economist (Free Exchange blog)* 1 (2013).

Westwood, S. J., S. Iyengar, S. Walgrave, R. Leonisio, L. Miller, and O. Strijbis, “The Tie That Divides: Cross-National Evidence of the Primacy of Partyism,” Unpublished paper (2015).

Williams, C. R., “Echo Chambers: Disagreement and Polarization in Bayesian Learning,” Working paper (2017).