Journal of Social and Political Psychology
jspp.psychopen.eu | 2195-3325

Original Research Reports

Moral Polarization and Out-Party Hostility in the US Political Context

Ben M. Tappin*[a], Ryan T. McKay[a]

[a] Department of Psychology, Royal Holloway University of London, Egham, United Kingdom.

Abstract

Affective polarization describes the phenomenon whereby people identifying as Republican or Democrat tend to view opposing partisans negatively and co-partisans positively. Though extensively studied, there remain important gaps in scholarly understanding of affective polarization. In particular, (a) how it relates to the distinct behavioural phenomena of in-party "love" vs. out-party hostility; and (b) to what extent it reflects a generalized evaluative disparity between partisans vs. a domain-specific disparity in evaluation. We report the results of an investigation that bears on both of these questions. Specifically, drawing on recent trends in political science and psychology, we hypothesize that moral polarization—the tendency to view opposing partisans' moral character negatively, and co-partisans' moral character positively—will be associated with behavioural hostility towards the out-party. We test this hypothesis in two preregistered studies comprising behavioural measures and large convenience samples of US partisans (combined N = 1354). Our results strike an optimistic chord: Taken together, they suggest that this association is probably small and somewhat tenuous. Though moral polarization itself was large—perhaps exceeding prior estimates of trait affective polarization—even the most morally polarized partisans appeared reluctant to engage in a mild form of out-party hostility. These findings converge with recent evidence that polarization—moral or otherwise—has yet to translate into the average US partisan wanting to express hostile and directly discriminatory behaviour toward their out-party counterparts.

Keywords: affective polarization, moral polarization, outgroup hostility, ingroup love, economic games

Journal of Social and Political Psychology, 2019, Vol. 7(1), 213–245, https://doi.org/10.5964/jspp.v7i1.1090

Received: 2018-11-05. Accepted: 2019-02-05. Published (VoR): 2019-03-28.

Handling Editor: Mark J. Brandt, Tilburg University, Tilburg, The Netherlands

*Corresponding author at: Department of Psychology, Royal Holloway University of London, TW20 0EX United Kingdom. E-mail: benmtappin@googlemail.com

Animosity between Republicans and Democrats is a salient feature of American political life. This animosity has been dubbed *affective polarization*; that is, the "tendency for people identifying as Republican or Democrat to view opposing partisans negatively and co-partisans positively" (Iyengar & Westwood, 2015, p. 691). Affective polarization is in evidence across a range of measures, and has been increasing over time (Iyengar, Sood, & Lelkes, 2012; Iyengar et al., 2018). For example, time series data from the American National Election Study indicate that the disparity in "warmth" that Democrats and Republicans (hereafter, *partisans*) express for their own party vs. the other party was greater in 2012 than at any point during the 34 preceding years; almost doubling in size since 1978 (Iyengar et al., 2012). Indeed, according to more recent analysis using this measure, the average partisan now feels almost three times more positive about the in-party than out-party (Iyengar et al., 2018).

Of course, affective polarization inferred from placement on these "feeling thermometers" says little about the behavioural manifestations and consequences of the phenomenon. In their recent review of affective polarization, Iyengar and colleagues (2018) cite and discuss evidence of such ramifications. In particular, US partisans are liable to allocate more money to the in-party than to the out-party in behavioural economic games (Carlin & Love, 2013; Iyengar & Westwood, 2015; Stagnaro, Dunham, & Rand, 2018); a pattern replicated cross-nationally in partisans in the United Kingdom, Belgium, Spain, South America and South Africa (Carlin & Love, 2018; Westwood, Iyengar, et al., 2018). In the US, furthermore, partisans are more likely to pursue online dating opportunities with politically similar others (Huber & Malhotra, 2017), and the magnitude of affective polarization positively correlates with avoidance of opposing partisans in a group problem-solving task (Lelkes & Westwood, 2017). Despite this evidence, Iyengar and colleagues (2018) note that it remains unclear (a) precisely how affective polarization relates to the distinct behavioural phenomena of "love" for the in-party vs. hostility towards the out-party (see also McConnell, Margalit, Malhotra, & Levendusky, 2018); and (b) to what extent affective polarization reflects a *generalized* evaluative disparity between partisans vs. a more domain-*specific* disparity in evaluation (e.g., that out-partisans are less trustworthy than in-partisans).

In this paper, we report the results of an investigation that bears on both of these questions. Specifically, we draw on recent trends in psychology and political science to hypothesize that *moral* polarization—that is, the tendency for people to view opposing partisans' moral character negatively, and co-partisans' moral character positively (cf. Iyengar & Westwood, 2015)—will be associated with behavioural expressions of out-party hostility. We test this hypothesis in two preregistered studies comprising behavioural economic game measures and large convenience samples of US partisans.

Group Identity and Moral Psychology in American Politics

Generally speaking, political (or ideological) conflict entails disagreement over which set of shared beliefs, values and practices make for a good and desirable society, and how this can be achieved (Jost et al., 2009). On this basis, even mild political disagreement is likely to be characterized by the belief that the in-party is more "moral" than the out-party; in other words, moral polarization. While the average American is not particularly committed to one ideological viewpoint over another, they do appear committed to a partisan *group identity*; that is, for the average American voter, politics may be more a case of Us vs. Them, than "our policy" vs. "their policy" (Kinder & Kalmoe, 2017). This can be expected to exacerbate moral polarization insofar as the human mind is primed to distinguish between ingroups and outgroups, and to interpret the social world in moral terms (Brewer, 1999; Haidt, 2012). Indeed, this proposition is consistent with the putative importance of both group identity and moral psychology in contemporary American politics.

Mason (2016, 2018), for example, documents that party identity in the US is increasingly in alignment with various other group identities, including race-, ideological-, and religious-based identities. Such "social sorting" may facilitate identification with the in-party and reduce the tempering influence of cross-cutting identities on out-party hostility (Mason & Wronski, 2018; Roccas & Brewer, 2002). At the same time, Ryan (2014) and Koleva and colleagues (2012) report evidence to suggest that moral psychological factors play an important and distinct role in the political preferences and behaviour of US partisans. In particular, the latter report that endorsement of a small number of "moral foundations" explains variance in attitudes across a wide range of US political issues—including gun control, immigration, equal marriage and abortion—beyond other relevant factors such as age, gender, ideology and interest in politics (Koleva et al., 2012). Ryan (2014), corroborating these results, finds that moral

conviction is common in partisans' policy attitudes—even for putatively *non*moral policy issues—and may undergird both political activism and political antagonism (see also Ryan, 2017; Skitka et al., 2005). Finally, recent work using data from Twitter suggests that posts about US political issues spread over the (ingroup) network to a greater extent if they contain *moral*-emotional language (vs. nonmoral-emotional language) (Brady et al., 2017).

In sum, the suffusion of group identities and moral psychology in contemporary American politics suggests that moral polarization—the tendency for people to view opposing partisans' moral character negatively, and co-partisans' moral character positively—may be particularly prominent among US partisans. We now consider the possible behavioural manifestations and consequences of moral polarization.

Moral Polarization and Out-Party Hostility

Though they may be conflated, ingroup "love" and outgroup hostility are distinct phenomena (Brewer, 1999; Lewis, Kandler, & Riemann, 2014). Whereas the former represents adulation for—and desire to help—members of one's own group, the latter represents hostility toward—and desire to harm—members of other groups. Indeed, people can exhibit ingroup love *without* exhibiting hostility towards a relevant outgroup (we note that the reverse case—outgroup hostility in the absence of ingroup love—seems less plausible a phenomenon). Where the two are appropriately disentangled, ingroup love appears to take psychological and behavioural primacy over outgroup hostility. That is, in intergroup conflict most people seem motivated primarily to benefit the ingroup, rather than to harm or discriminate against the outgroup (Brewer, 1999; Halevy et al., 2008; Weisel & Böhm, 2015).

The primacy of in-party love over out-party hostility is evident in US political conflict in particular. For example, McConnell and colleagues (2018) recently studied partisans' willingness to complete paid work for an employer. Randomizing the political identity of the employer, they found that relative to a control group partisans demanded a lower price to complete additional work for an in-party employer (in-party love), but they did not demand a higher price for an out-party employer (i.e., no evidence of out-party hostility). The authors found a similar pattern of results in another study. Across a series of five studies, Lelkes and Westwood (2017) observed that affective polarization was positively related to more lenient sanctions for in-party agents—ingroup love—but was unrelated to preferences over sanctions for out-party agents. The authors conclude that "affective polarization, like other in-group–out-group divides documented in psychology, is more about in-group love than out-group hate." (p. 496.)

However, in some contexts and for some people hostility towards the outgroup is also clearly an important motivation. For example, one study found that monetary donations in the Dictator Game were lower to an opposing partisan (relative to control) by roughly the same magnitude as they were higher to a co-partisan (Iyengar & Westwood, 2015, Study 4). In other words, evidence of *both* in-party love and out-party hostility. Carlin and Love (2018) likewise observed evidence of both in-party love and out-party hostility in the domain of partisan trust behaviour. Recent work suggests that displays of out-party hostility are more common when group identities are defined—and the relevant groups divided—along *morality*-based lines (Parker & Janoff-Bulman, 2013; Weisel & Böhm, 2015). For example, using a novel behavioural economic game with subjects in Germany, Weisel and Böhm (2015) found that game decisions indicative of hostility toward the outgroup were more common towards supporters of the National Democratic Party (NPD)—considered neo-Nazi and widely morally opposed—than towards supporters of other political parties in Germany. While the NPD are arguably unique in their moral, cultural and historical significance in Germany, this result suggests that variance in behavioral expressions of out-party hostility may be explained by variance in moral polarization in the United States. That is, as the perceived moral "gap" between in- and out-party widens, behavioural expressions of out-party hostility increase in likelihood. This

could help explain previous mixed findings of in-party love vs. out-party hostility in the US political context (Carlin & Love, 2018; Iyengar & Westwood, 2015; Lelkes & Westwood, 2017; McConnell et al., 2018).

The logic of this hypothesis is further supported by analyses of real-world ideological conflict. Analyses of the patterns of thinking in militant extremism (Giner-Sorolla et al., 2012; Saucier et al., 2009), violent political and religious conflict (Ginges et al., 2011), and genocidal regimes (Koonz, 2003; Reicher et al., 2008) identify the *extolling of ingroup virtue* as a persistent theme. To eschew self-interest and contribute to ingroup ends, would-be contributors must feel sufficiently persuaded of the moral righteousness of their comrades, and of their cause. When the ingroup and its cause are perceived as just, the personal costs involved in intergroup conflict may be tolerable or even desirable (Saucier et al., 2009). *Moral demonization of the outgroup* is another recurring theme in real-world ideological conflict (Giner-Sorolla et al., 2012; Halperin, 2008; Reicher et al., 2008; Saucier et al., 2009). In genocides, for example, propaganda depicting outgroup targets as nefarious agents with hostile intentions is reputedly commonplace, and is thought to be a deliberate strategy to rally public support for genocidal policy (Bilewicz & Vollhardt, 2012). Morally demonized outgroups may be perceived as an existential threat to the ingroup (Giner-Sorolla et al., 2012), and, in contexts where the latter is morally championed, this feeds a compelling Manichean survival narrative of good against evil (Reicher et al., 2008; Saucier et al., 2009). Under such conditions, expressions of outgroup hostility may become morally mandated (Skitka & Mullen, 2002). Furthermore, given the close tie between moral cognition on the one hand, and peoples' worldviews and ideologies on the other, our hypothesis is also corroborated by work showing that prejudice is more likely towards those who represent a threat to the latter (e.g., Brandt, 2017; Crawford, Brandt, Inbar, Chambers, & Motyl, 2017).

Overview of Studies

Following the rationale outlined above, we hypothesized that moral polarization would be associated with behavioural expressions of out-party hostility in the US political context. We tested this hypothesis in two preregistered studies—an initial study and a direct replication—comprising large convenience samples of US partisans, and a behavioural economic game measure of outgroup hostility.

Methods

Both studies were preregistered on *AsPredicted*: https://aspredicted.org/e3hw9.pdf (link to Study 1 protocol); https://aspredicted.org/tiuw7.pdf (Study 2 protocol). To avoid unnecessary repetition—Study 2 was a close replication of Study 1—we present the methods and results of the studies together.

Samples

We sought to recruit 450 subjects in Study 1 and 900 subjects in Study 2. Subjects were supporters of the US Republican or Democratic Party, recruited via Amazon's Mechanical Turk (MTurk), an online labour market commonly used for psychological research (Arechar et al., 2018; Chandler & Shapiro, 2016; Rand, 2012). Subjects recruited on MTurk cannot be considered demographically representative of the wider US population; for example, they are more educated and more liberal/Democrat, among other demographic differences (Chandler & Shapiro, 2016). Despite this, there is evidence that political partisans recruited via MTurk are *psychologically* similar to partisans in nationally representative samples of US adults (Clifford et al., 2015). In particular, they score similarly on measures of personality and values related to political ideology (*ibid.*). Therefore, while there are documented

constraints on the generalizability of results obtained from MTurk samples, the results of Clifford and colleagues (2015) suggest that subjects recruited via MTurk are not psychologically *incomparable* to the average US partisan.

The sample size for Study 1 was determined by power analysis (Faul et al., 2009), according to which we required N = 391 to detect an odds ratio of 1.4 in our primary binomial logistic regression analysis (key parameters: Two-tailed test; α = .05; power = 0.9; Pr(Y=1|X=1) H0 = 0.5). We oversampled by approximately 15% to guard against power loss due to planned data exclusions. Sample size after data collection was N_{S1} = 454 in Study 1 (52.86% female, M_{age} = 36.73, SD_{age} = 12.64). Slight oversampling is the result of subjects not submitting their completion code on MTurk despite completing the study (i.e., meaning additional subjects were able to complete the study). In Study 2, we doubled the target sample size of Study 1 and recruited N_{S2} = 900 (59.33% female, M_{age} = 36.28, SD_{age} = 11.40).

Measures

Out-Party Hostility

Intergroup conflict behaviours typically impose personal costs on the individuals involved. For example, in the risk of harm or injury, opportunity costs, and physical exertion. Why do individuals pay these costs? Halevy and colleagues (2008, p. 405) distinguish between two motivations: the "altruistic desire" to help the ingroup (i.e., ingroup love) vs. the "aggressive drive" to hurt the outgroup (i.e., outgroup hostility). Of course, hostility towards an outgroup may ultimately be conceived of as ingroup love; specifically, by increasing the ingroup's relative advantage over the outgroup (Tajfel et al., 1971). Nevertheless, identifying the conditions under which people are willing to exhibit behavioural hostility towards an outgroup (vs. not) is important irrespective of whether or not that hostility is ultimately borne of ingroup love. Our hypothesis is that moral polarization constitutes one such condition.

To distinguish behavioural expressions of ingroup love from those of outgroup hostility, Halevy and colleagues (2008) designed the Intergroup Prisoner's Dilemma-Maximizing Difference (IPD-MD) game. We use an adapted version of the IPD-MD—its positive variant (Weisel & Böhm, 2015)—in our design. In our positive variant of the IPD-MD, subjects are assigned to a subgroup with two supporters of the same political party; that is, the *in-party*. This subgroup is matched with another subgroup of three supporters of the opposite party; the *out-party*—forming a collective group of six players in total.

Each subject is faced with three choices about how to allocate money. Option #1 provides US$5 only to the focal subject—the self-interested choice. Option #2 provides $2.50 to the each of the ingroup players but nothing to any other players in the game. Option #3 provides $2.50 to each player in the game. The decision options (as they were shown to Democratic Party subjects) are displayed in Figure 1. The decision option relevant to our hypothesis is Option #2. This decision option evinces a willingness to pay a personal cost to benefit the in-party (i.e., forsaking Option #1; self-interest), while simultaneously refusing to benefit members of the out-party at no extra cost to oneself or to members of one's in-party (forsaking Option #3; the collective interest). Following Halevy et al. (2008) and Weisel and Böhm (2015), we thus interpret decision Option #2 as an expression of out-party hostility[i].

This version of the IPD-MD is called "positive variant" because outgroup hostility is characterized as denial of (costless) help rather than subtraction of existing money from the outgroup—as in the original IPD-MD. We decided to use the positive variant here to maximize variance in the Option #2 choice, because past work finds that very few individuals choose Option #2 when it involves taking money away from the outgroup (Halevy et al., 2008;

Weisel & Böhm, 2015). Moreover, choosing Option #2 in the positive variant of the game is strongly correlated with the same choice in the negative variant (r_{range} = [.48, .72], see Weisel, 2015); supporting an inference of conceptual similarity.

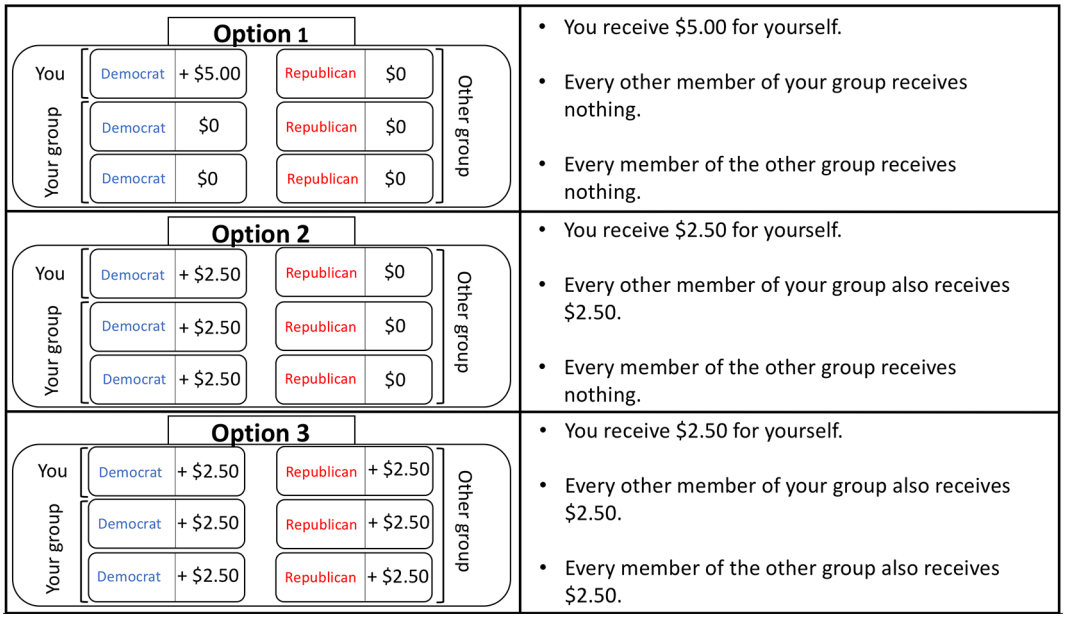

Figure 1. Decision options in the positive variant of the intergroup prisoner's dilemma maximizing-difference game used in Studies 1 and 2.

Note. Self-identified Democrats saw the displayed decision option screen. For subjects who identified as Republicans, the Republican and Democrat labels were reversed (i.e., Republicans were indicated as "your" group, and Democrats were indicated as the "other" group).

Moral Polarization

To measure this variable, subjects completed a trait judgment task. Each subject was asked to judge the extent to which 5 positive and 5 negative moral traits described each of two targets: (i) the "average Democratic Party voter" and (ii) the "average Republican Party voter". Subjects also rated the social desirability of each trait. All trait ratings were provided on a 1-7-point scale. The target ratings were anchored from "Not at all" to "Very much so"; the desirability ratings were anchored "Very undesirable" to "Very desirable". The traits comprised personality descriptors such as *trustworthy*, *fair*, *manipulative* and *prejudiced*, and were embedded alongside a mix of 20 nonmoral traits. Table 1 displays the full list of traits used in Studies 1 and 2.

The traits used in Study 1 were taken from prior work (Goodwin et al., 2014; Tappin & McKay, 2017) and were chosen to represent three distinct domains of social perception: morality, agency and sociability (Leach et al., 2007). In a large trait-norming study, Goodwin and colleagues (2014, Study 1) asked N = 1084 respondents how useful each of 170 traits were in providing information about higher-level person characteristics—such as "ability" or "morality". Tappin and McKay (2017) averaged across these ratings to create composite scores indicating how well each trait corresponded to the domains of morality ("morality/immorality", "character"), agency ("ability", "agency") and sociability ("warmth", "communion"). They then selected 10 traits from each domain that (i) scored highly on the focal domain, and (ii) scored as low as possible in the other two domains. We adopted the traits

from Tappin and McKay (2017) for Study 1 here. The traits we used in Study 2 were slightly modified. Specifically, we replaced several traits with new traits from the aforementioned dataset of normed trait adjectives (Goodwin et al., 2014, Study 1). We did this to minimize any residual overlap between the three trait domains.

Table 1

Traits Used in Studies 1 and 2

Trait domain	Positive traits	Negative traits
Morality		
	Honest	Insincere
	Trustworthy	Prejudiced
	Fair	Disloyal
	Respectful (Just)	Manipulative (Violent)
	Principled	Deceptive (Greedy)
Agency		
	Hardworking (Intelligent)	Lazy
	Knowledgeable	Undedicated (Incompetent)
	Competent (Organized)	Unintelligent (Unproductive)
	Creative	Unmotivated
	Determined	Illogical (Weak)
Sociability		
	Sociable	Cold (Negative)
	Cooperative (Playful)	Disagreeable
	Warm (Happy)	Rude (Reckless)
	Family-orientated (Funny)	Humorless
	Easygoing	Uptight

Note. Traits outside of parentheses are used in Study 1. Traits inside parentheses replaced the preceding traits in Study 2.

We calculated subjects' moral evaluation of each target (average Democratic and Republican party voter) in two different ways; ultimately, resulting in two different indices of moral polarization. Our preregistered index of moral polarization is outlined immediately below, followed by description of the additional index that was not preregistered. We computed the latter index in order to maximize the validity of our inferences about the association between moral polarization and out-party hostility[ii].

Correlation (Preregistered) Index. Subjects' moral evaluation of each target was computed as the correlation between (i) their social desirability ratings for the moral traits, and (ii) their Democratic/Republican target ratings for the moral traits. Thus, each subject had two "coefficients of moral evaluation", describing the extent to which they ascribed desirable and undesirable moral traits to each target. Positive coefficient values indicate that the ascription of moral traits to the target *positively* correlated with the perceived desirability of those traits. In contrast, therefore, negative coefficient values indicate that the ascription of moral traits *negatively* correlated with the desirability of the traits. Because each subject rated the desirability of each trait, the coefficient values—representing subjects' moral evaluation of the targets—are sensitive to subjects' *idiosyncratic* beliefs about the desirability of the moral traits. This has the advantage of allowing for individual differences in which moral traits people consider more vs. less desirable when computing their coefficients of moral evaluation for each target. This is important because

previous work suggests foundational differences in the moral preferences of Democrats (or "liberals") and Republicans (or "conservatives") (e.g., Graham et al., 2009).

Finally, whether the subject identified as Democrat or Republican informed which coefficient represented moral evaluation of the in-party ($r_{inParty}$) and out-party ($r_{outParty}$). For example, for a self-identified supporter of the Republican Party, $r_{inParty}$ corresponded to the coefficient of moral evaluation for the Republican target, and $r_{outParty}$ for the Democratic target (and vice versa for subjects who identified as supporters of the Democratic Party). The difference between these coefficients of moral evaluation ($r_{inParty} - r_{outParty}$) was taken as the discrepancy in moral evaluation between the in-party and out-party; that is, the preregistered measure of moral polarization[iii]. A weakness of this index of moral polarization, however, is that the correlation coefficients are sensitive only to the *relative*—not absolute—values of desirability judgments and target trait ascriptions provided by subjects. One consequence of this is that the index cannot discriminate between subjects whose patterns of target trait ascriptions are at the *extreme* ends of the scale vs. just past the midpoint of the scale, as long as those ascriptions have the same *relative* order (holding desirability judgments constant). The correlation index may thus obscure important *absolute* differences in subjects' moral polarization. Furthermore, the correlation index of moral polarization is necessarily bounded between -2 and +2 (because it is the sum of two correlation coefficients, one for the in-party and one for the out-party); possibly skewing the values. For these reasons, we computed a second index of moral polarization—the weighted-sum index—which is superior because it is sensitive to both (i) the idiosyncratic desirability judgments of subjects *and* (ii) the absolute values of target trait ascriptions. It is also not bounded between values of -2 and +2.

Weighted-Sum Index. This index assigns weights to each target trait ascription. The weights correspond to how desirable/undesirable the trait is judged to be by the subject. Specifically, a desirability judgment of "1", indicating that the trait is "extremely undesirable", corresponds to a weight of -1; whereas a judgment of "7" (extremely *desirable*) corresponds to a weight of +1. Judgment values in-between 1 and 7 (i.e., 2-6) correspond to weights of -0.67, -0.34, 0, +0.34 and +0.67 (rounded), respectively. We multiply each raw trait ascription by these weights—meaning the resultant trait ascription is weighted *proportional to its perceived desirability*. For example, assume a trait ascription judgment of "7" for the trait "honest" (indicating strong ascription of that trait to the target). Further assume the subject's desirability judgment is "6" for this trait. The weighted trait ascription is thus 7 * 0.67 ≈ 4.67. Suppose a different subject's desirability judgment is "7" for this trait (but their raw trait ascription is the same), then their weighted trait ascription is 7 * 1 = 7.

After weighting all trait ascriptions by their respective desirability judgments according to this method, for each subject we then summed across their weighted trait ascriptions. This provides two weighted-sum scores per subject: one corresponding to the in-party, and the other corresponding to the out-party. Higher values denote relatively more positive moral evaluation of the target. As before, in a final step, for each subject we subtract the out-party weighted-sum score from the in-party weighted-sum score to give the index of moral polarization (the possible range of values is thus -60 to +60).

Other Variables

We collected additional variables after the behavioural economic game and trait rating task. These variables were collected for the purpose of secondary preregistered and exploratory analyses. First, we asked each subject which of the three decision options they believed their two in-party members, and three out-party members, had chosen. Second, we asked subjects to rate the extent to which they believed that their out-party (i) threatened the "power,

resources, or safety of the US and its citizens", and (ii) threatened the "values or identity of the US and its citizens" (Stephan et al., 2011). Lastly, we asked subjects to rate (iii) the extent to which they believe the Democratic and Republican Party are in "direct competition". Ratings for (i), (ii), and (iii) were provided on 7-point Likert scales, anchored from 1 = "Not at all" to 7 = "Very much so".

Procedure

The procedure in both studies was substantively identical and we recruited unique samples in each (i.e., subjects who took part in Study 1 were prevented from taking part in Study 2). All subjects provided informed consent, before completing a brief screening questionnaire. This questionnaire identified whether the subject was a supporter of the Democratic Party or the Republican Party, and included other demographic questions such as age, gender, religious affiliation, and ethnicity. Importantly, subjects were not made aware of the specific purpose of the screening questionnaire (to minimize false responding). Subjects who identified with either the Democratic Party or Republican Party were eligible to continue with the study, whereas supporters of a political party other than these (including "none") were directed to an end-of-study message and were unable to continue. The Study 1 sample was skewed Democrat (Study 1 = 67.62% Democrat). In Study 2, we balanced the number of Democrats and Republicans by recruiting approximately equal numbers of each (Study 2 = 50.11% Democrat).

Eligible subjects then completed the trait judgment task. They judged the extent to which each of 30 traits (see Table 1) described (i) the "average Democratic Party voter" and (ii) the "average Republican Party voter". They also rated (iii) the social desirability of the traits. Subjects rated all 30 traits according to either (i), (ii), or (iii), before moving onto the next set of ratings, and the order of these three sets of judgments was counterbalanced across subjects. The presentation order of the traits themselves was randomized across each rating set and subject.

Following this task, subjects took part in the economic game. They read instructions detailing the structure of the game and were shown an example set of decisions (and the resultant pay offs). Those who identified as Republican were presented with instructions specifying two other Republicans as their subgroup members (and three Democrats as members of the other subgroup), and vice versa for Democrats. After these instructions, subjects made their decision about which option to choose (i.e., Option 1, 2, or 3). We informed them that six individual decisions (three from Democrats, three from Republicans) would be combined, and the calculated bonuses paid out to one group of six—selected at random—after the survey had ended (which was true, there was no deception). After making their own decision, each subject indicated which decision they believed each of the other players had chosen, and responded to the threat and competition questions described above. Finally, at the end of the study, subjects were asked whether they had adequately understood the economic game before making their decision (yes/no), and they provided feedback on the study. In addition to any bonuses, all subjects were paid a base fee of $1 for taking part.

Results

All analyses were conducted in the R environment (v. 3.4.0, R Core Team, 2017), using R Studio (v. 1.1.423, RStudio Team, 2016). The R packages used in data analysis were: *scales* (v. 1.0.0, Wickham, 2018), *coin* (v. 1.2-2, Hothorn et al., 2008), *gridExtra* (v. 2.3, Auguie, 2017), *ggthemes* (v. 3.4.0, Arnold, 2017), *dplyr* (v. 0.7.7, Wickham et al., 2018), *ggplot2* (v. 3.0.0, Wickham, 2016), *reshape* (v. 0.8.7, Wickham, 2007), *plyr* (v. 1.8.4,

Wickham, 2011), *metafor* (v. 2.0-0, Viechtbauer, 2010), *datatable* (v. 1.10.4-3, Dowle & Srinivasan, 2017) and *psych* (v. 1.8.12, Revelle, 2018). The raw data and analysis scripts to reproduce the results and figures reported in this paper are available online via the project hub on the Open Science Framework: https://osf.io/mceqh/.

Analytic Strategy

The results section proceeds as follows. First, we report descriptive statistics regarding the key variables of interest: (i) choices in the IPD-MD and (ii) moral polarization. Second, we outline a series of preregistered data exclusions that were implemented prior to our primary analyses. Third, we report our primary preregistered analyses—which consist in regressing the correlation index of moral polarization on choice made in the IPD-MD. Fourth, we present a series of exploratory analyses that investigate the robustness of the association between moral polarization and out-party hostility observed in the preregistered tests. These analyses begin by regressing the alternative, weighted-sum index of moral polarization on choice in the IPD-MD. Because this index is superior to the preregistered moral polarization index (see Methods)—and produces a qualitatively similar result as the primary analyses—the subsequent exploratory analyses are conducted using the weighted-sum index only.

The second key exploratory test that we perform is on the *interaction* between (i) in-party moral evaluation and (ii) out-party moral evaluation variables in predicting out-party hostility. Recall that our index of moral polarization is a *single* variable, comprised of (i) and (ii) and computed as: in-party score – out-party score. Combining the variables in this way is faithful to the operationalization of affective polarization (e.g., Iyengar et al., 2012; Lelkes & Westwood, 2017). However, it obscures whether the conjunction of in- and out-party moral evaluation *per se* predicts outcomes, or whether outcomes are predicted mostly by one of the variables only (i.e., *either* in-party *or* out-party evaluation). This obscurity is particularly important here, given that our hypothesis about subjects' behaviour in the IPD-MD speaks to the former—that is, the conjunction of in- and out-party moral evaluation—not the latter (main effects of in-party *and/or* out-party evaluation).

To provide further intuition for why this is the case, consider the three choices faced by subjects in the IPD-MD (Figure 1). Those who morally champion the in-party are most likely to forsake self-interest (Option #1) to help their fellow partisans—narrowing their choice options to #2 and #3. If they are ambivalent about the morality of the out-party then they may have little motivation not to help them, too—thus choosing Option #3. On the other hand, if they both morally champion the in-party *and* morally demonize the out-party, then Option #2—helping the in-party but denying the out-party—becomes more attractive. Thus, a particular *conjunction* of in-party moral evaluation (max.) and out-party moral evaluation (min.) ought to best predict choice of Option #2 (according to our hypothesis). To put it another way: out-party moral evaluation is only likely to predict choice of Option #2 (vs. #3) when in-party moral evaluation is strong positive; when in-party moral evaluation is negative, all subjects would presumably choose Option #1 however they feel about the out-party. The best test of our hypothesis is therefore on the interaction between these variables—not the single index analysis. However, because affective polarization is conceptualized and operationalized as a single variable in prior relevant work (Iyengar et al., 2012; Lelkes & Westwood, 2017), and because we preregistered the single-index analysis, we report both the single-index and interaction tests—while noting that our hypothesis is best evaluated by the latter.

Finally, we report several additional exploratory tests that bear on extant work in social and political psychology. We also note that, because Study 2 was a close replication of Study 1, after reporting each of the study-specific effect size estimates in the primary and sensitivity analyses, we also report the associated *meta-analytic* estimate

(i.e., computed across studies). All meta-analytic estimates are fixed effects and the meta-analyses were not preregistered.

Descriptive Statistics

Table 2 displays the frequency and corresponding percentages of choices made in the IPD-MD game in Studies 1 and 2 (the full samples are displayed, before any data exclusions).

Table 2

Economic Game Decisions in Studies 1 and 2

Subject Political Affiliation	Self-Interest (Option 1)	Out-Party Hostility (Option 2)	Collective Interest (Option 3)	Total
Study 1				
Republican	40 (27.2%)	22 (15.0%)	85 (57.8%)	147 (100%)
Democrat	72 (23.6%)	54 (17.7%)	179 (58.7%)	305 (100%)
Total	112 (24.8%)	76 (16.8%)	264 (58.4%)	452 (100%)
Study 2				
Republican	113 (25.3%)	72 (16.1%)	262 (58.6%)	447 (100%)
Democrat	131 (29.6%)	57 (12.9%)	255 (57.6%)	443 (100%)
Total	244 (27.4%)	129 (14.5%)	517 (58.1%)	890 (100%)

Note. The numbers outside parentheses are frequencies and the numbers inside parentheses are row-wise percentages. $N = 2$ observations are missing from Study 1 and $N = 10$ observations from Study 2 due to missing values for choice decision in the economic game.

Table 3 displays the median values of the coefficients of moral evaluation (i.e., the constituent variables of the correlation index of moral polarization). In particular, displayed are the median coefficients pertaining to moral evaluation of the Democrat and Republican targets, separately for Democratic- and Republican-identifying subjects. Also displayed are the median coefficient values pertaining to in-party and out-party targets; in other words, collapsing *across* Democratic and Republican targets/subjects (as described in the Methods). We compare the coefficients using Wilcoxon signed-rank tests. The resultant test values in Table 3 reveal a robust discrepancy in moral evaluation for the in-party vs. out-party—convincing evidence of moral polarization—among both Democratic- and Republican-identifying subjects. This is confirmed by visualizing the distribution of the correlation index of moral polarization (recall computed as $r_{\text{inParty}} - r_{\text{outParty}}$), displayed in Figure 2. Scores greater than zero imply more positive moral evaluation of the in-party than out-party. The distribution of the exploratory weighted-sum index of moral polarization is also displayed in Figure 2, and it shows a qualitatively similar pattern. Indeed, the two indices of moral polarization (correlational vs. weighted-sum) are strongly correlated, $r_{S1}(440) = .84$, $p < .001$, 95% CI [.81, .86]; $r_{S2}(872) = .86$, $p < .001$ [.84, .88].

Table 3

Median Coefficients of Moral Evaluation in Studies 1 and 2

Subject Political Affiliation	Moral Evaluation					
	Dem. Target	Rep. Target	(Pseudo) Difference	In-Party Target	Out-Party Target	(Pseudo) Difference
Study 1						
Republican	-.34	.89	-0.94***			
Democrat	.88	-.50	1.10***			
Combined				.88	-.46	1.04***
Study 2						
Republican	-.34	.85	-0.91***			
Democrat	.90	-.42	1.08***			
Combined				.88	-.37	0.99***

Note. Values for targets are the median correlations between ratings of trait desirability and trait ascription. The pseudo-difference is computed by Wilcoxon signed-rank test. Study 1 N = 442 (N = 12 subjects could not be included due to uniform responding on the trait judgment task); Study 2 N = 874 (N = 26 subjects were not included due to uniform responding and/or missing values on the trait judgment task).
*p < .05. **p < .01. ***p < .001.

Data Exclusions

As specified in the preregistered protocols, for the primary analysis we excluded subjects who fulfilled one or more of several criteria. First, we excluded those who failed one or more of three attention checks that were embedded in the trait judgment task, N_{S1} = 23 (5.07%); N_{S2} = 28 (3.11%). Second, those who provided incomplete data in the trait judgment task, N_{S1} = 1 (0.22%); N_{S2} = 7 (0.78%), or IPD-MD game, N_{S1} = 2 (0.44%); N_{S2} = 10 (1.11%). Third, those who clicked through the IPD-MD game instructions too quickly to read them; defined as a recorded page submission time of less than 10 seconds on one or more of three instructions pages, N_{S1} = 74 (16.30%); N_{S2} = 175 (19.44%). Fourth, those who reported that they did not understand the IPD-MD game instructions, N_{S2} = 8 (1.76%); N_{S2} = 17 (1.89%). Fifth, and finally, those subjects who responded *uniformly* on the trait judgment task—that is, recorded zero variance for any type of moral trait judgment (i.e., ratings for the Democratic target, Republican target, and/or social desirability), N_{S1} = 12 (2.64%); N_{S2} = 22 (2.44%); this was necessary because a lack of variance prevents correlation coefficients—required for the key measure of moral polarization (see Methods)—from being computed.

In addition to these preregistered exclusion criteria, we identified and excluded duplicate responses (i.e., multiple responses from the same subject) via subjects' unique MTurk IDs, N_{S1} = 1 (0.22%); N_{S2} = 26 (2.89%). After all data exclusions, we thus retained N_{S1} = 354 and N_{S2} = 671 for the primary preregistered analyses.

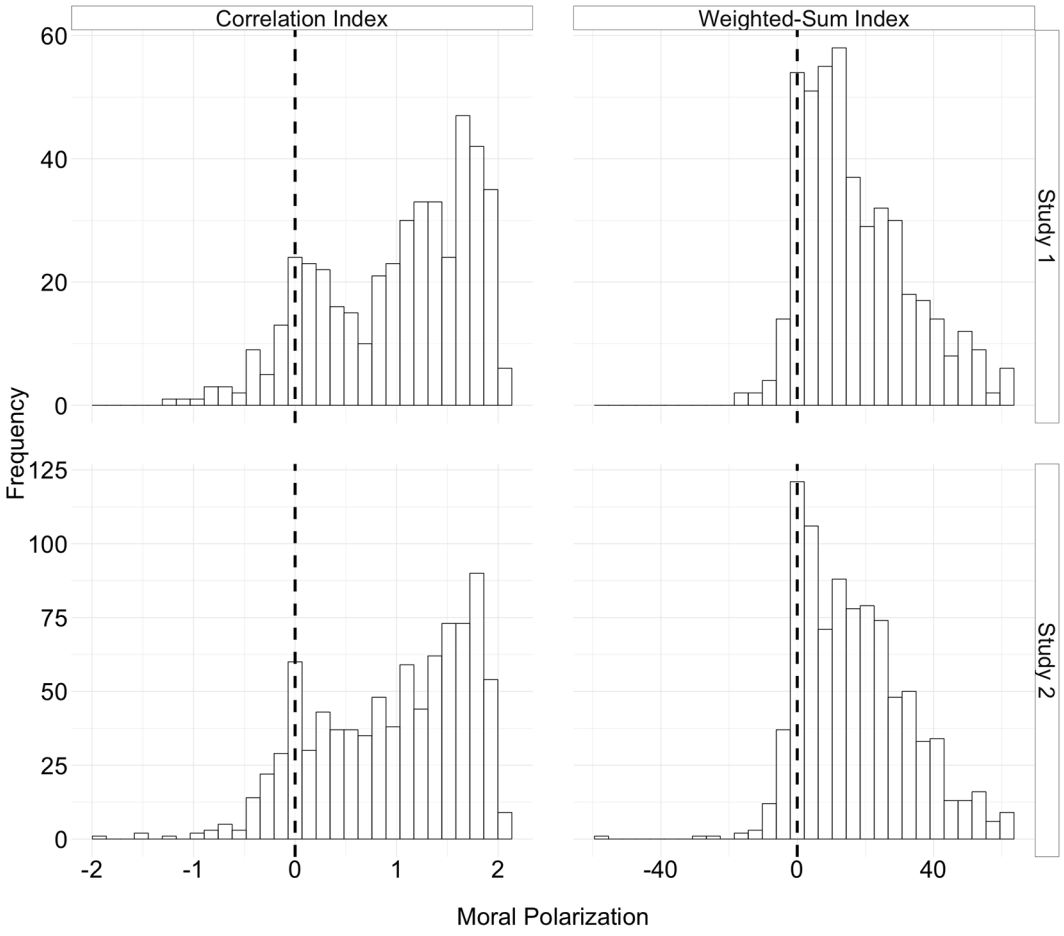

Figure 2. Distribution of moral polarization values according to each index in Studies 1 and 2.

Note. Each subject has a single value in each panel. Correlation Index: Study 1 N = 442, Study 2 N = 874; Weighted-Sum Index: Study 1 N = 454, Study 2 N = 896.

Primary Preregistered Analyses

We fitted binomial logistic regression models to the data. Recall that the outcome of interest is choosing Option #2—expression of out-party hostility—in the IPD-MD game (dummy-coded). The predictor variable is the correlation index of moral polarization, computed as the difference between subjects' coefficient of moral evaluation for the in-party and that for the out-party ($r_{\text{inParty}} - r_{\text{outParty}}$). Higher values thus correspond to relatively greater moral polarization.

Moral polarization was positively associated with out-party hostility in the IPD-MD game, in both studies: Odds Ratio$_{S1}$ (OR$_{S1}$) = 1.73, p = .027, 95% CI [1.06, 2.81]; OR$_{S2}$ = 1.51, p = .025 [1.05, 2.17]. These odds ratios are plotted in Figure 3, indexed by *Primary Preregistered* on the y-axis. The meta-analytic OR was 1.59, p = .002 [1.19, 2.12]. According to the models, subjects at the upper limit of moral polarization—that is, a value of 2 (indicating a coefficient value of +1 for the in-party and -1 for the out-party)—had a predicted probability of 0.21 (Study 1) and 0.17 (Study 2) of expressing out-party hostility, respectively. In contrast, subjects whose moral evaluation of the in-party and out-party were similar—a score of 0 on the correlation index of moral polarization (indicating

no difference in coefficient values for the in-party and out-party)—had a predicted probability of 0.08 of expressing out-party hostility (in both Studies 1 and 2).

Exploratory Analyses

We conducted a series of preregistered and exploratory analyses to examine the robustness of the above result. These are reported below.

Weighted-Sum Index of Moral Polarization

Before fitting the models with the alternative weighted-sum index of moral polarization, we rescaled this index to lie between -2 and +2 to facilitate comparison of the ORs with the primary preregistered models. As displayed in Figure 3, the ORs for the weighted-sum index were statistically significant and very similar in size to the ORs in the primary preregistered models. Indeed, the meta-analytic OR for the weighted-sum index of moral polarization predicting out-party hostility was near identical to that for the preregistered index; OR = 1.60, $p < .001$ [1.22, 2.10].

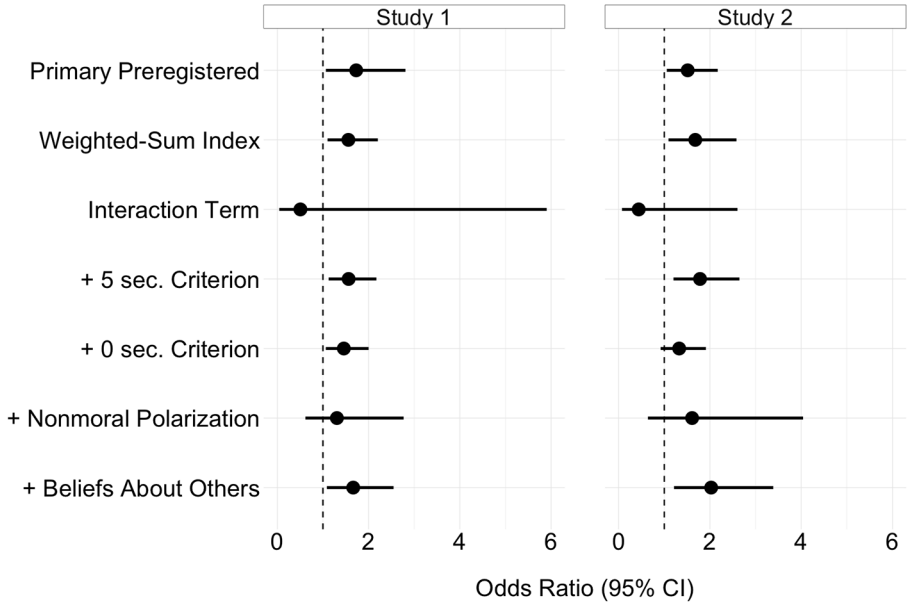

Figure 3. Results of preregistered and exploratory sensitivity analyses in Studies 1 and 2.

Note. Panels display Odds Ratios (with 95% confidence intervals) from different models with moral polarization predicting out-party hostility. Models are indexed on the y-axis. The + sign indicates that these models are versions of the Weighted-Sum Index model.

Interaction Test

Next, we fitted binomial logistic regression models to the data and specified the predictor variables as (i) in-party moral evaluation, (ii) out-party moral evaluation and (iii) the interaction between (i) and (ii). As shown in Figure 3 (*Interaction Term*), both interaction estimates had wide confidence intervals and neither was statistically significant.

What explains the disparity between this result and the results from the models with the single index of moral polarization? Dropping the interaction term from the interaction models reveals the reason. In Study 1, there was a main effect of in-party moral evaluation on probability of expressing out-party hostility such that a more positive

evaluation was associated with an increased probability (OR_{S1} = 3.25, p = .014 95% CI [1.27, 8.35]); whereas, in Study 2, there was a main effect of *out*-party moral evaluation, such that a less positive evaluation was associated with an increased probability of out-party hostility (OR_{S2} = 0.45, p = .029 [0.22, 0.92]). In other words, the results of the single index analyses in Studies 1 and 2 were driven by *main effects* of in-party evaluation and out-party evaluation, respectively (the opposite main effect in each of the models did not significantly improve model fit). These results indicate that the conjunction of (i) moral championing of the in-party and (ii) moral demonization of the out-party does not uniquely predict out-party hostility—that is, contrary to our hypothesis. Instead, out-party hostility was associated with in-party (Study 1) or out-party (Study 2) moral evaluation *per se*.

Instructions Page Exclusion Criterion

As reported in the data exclusions subsection, the number of subjects excluded for clicking through one or more of the IPD-MD game instructions too quickly (< 10 seconds) was relatively high in both studies. We therefore repeated the weighted-sum single-index analyses after implementing a more conservative exclusion criterion. Specifically, in one exploratory analysis we reduced this exclusion criterion to < 5 seconds (i.e., *5 sec. Criterion* models), and, in another, we removed this particular criterion altogether (*0 sec. Criterion* models). As plotted in Figure 3, in the former case (*5 sec. Criterion*) the ORs for the single-index moral polarization variable remained similar in size and statistically significant. The meta-analytic OR was 1.65, p < .001 [1.28, 2.12]. In the *0 sec. Criterion* models, in contrast to Study 1, the OR decreased noticeably in size and was no longer statistically significant (p > .05) in Study 2. The meta-analytic OR was 1.40, p = .006 [1.10, 1.78]. Overall, we conclude that the single-index moral polarization result is robust to more conservative specifications of the instructions page exclusion criterion.

Nonmoral Polarization

Recall that subjects also rated *non*moral traits in the trait judgment task, corresponding to the domains of agency and sociability. We preregistered our intention to investigate whether (single-index) moral polarization was associated with out-party hostility independent of polarization in these nonmoral domains of evaluation. We thus computed polarization scores for traits in the agency and sociability domains—in the same fashion as the weighted-sum index of moral polarization variable was computed—and entered these new variables as additional predictors in the weighted-sum single-index models. As can be seen in Figure 3 (*Nonmoral Polarization*), the confidence intervals on the ORs for moral polarization increased in size after modelling the two nonmoral polarization variables. Thus, the ORs were no longer statistically significant (at p < .05) in either study. The meta-analytic OR was 1.42, p = .238 [0.79, 2.54] (we note that the agency/sociability polarization predictors were not statistically significant predictors of out-party hostility in either model). This raises the question of whether moral polarization is distinct from nonmoral polarization in the data. We examine this question in detail in the next section.

Moral vs. Nonmoral Polarization

We sought to compare the magnitude of moral polarization to the magnitude of *non*moral polarization among partisans. We did this in two ways. First, we computed the mean rating given on each individual trait (i.e., for each trait in Table 1) as they were ascribed to each target (in-party, out-party). These mean ratings are plotted in Figure 4 as a function of the trait domain (agency, morality, sociability) and trait valence (negative, positive) (denoted by the faded small data points). We also plot the subsequent mean computed over these individual trait rating means (denoted by the solid large data points). As can be seen in the figure, across all trait domains, when rating the *in*-party target subjects ascribed *positive* traits more strongly than negative traits. However, across studies this

valence gap appeared to be slightly larger in the agency and morality domains vs. the sociability domain. When rating the *out*-party target, in contrast, subjects tended to ascribe *negative* traits more strongly than positive traits; but only in the morality and sociability domains (in the agency domain, a similar valence gap was not evident).

Figure 4. Mean trait ratings as a function of party target (in-party, out-party), trait valence (negative, positive) and trait domain (agency, morality, sociability) in Studies 1 and 2.

Note. The faded small data points denote the mean rating for the individual traits in each valence/domain category (see Table 1 for the traits). The data points are horizontally jittered to aid visibility. Each individual trait mean is computed over N = 453 subjects in Study 1 and over N = 874 in Study 2. The solid large data points denote the mean computed over the individual trait rating means. Error bars are 95% CI.

To formally compare the domain-specific magnitudes of polarization, we conducted Wilcoxon signed-rank tests between the weighted-sum indices of polarization corresponding to each of the three trait domains. That is, we compared the weighted-sum (single-) index of moral polarization with the corresponding indices of polarization on the agency and sociability traits. Recall that higher values on all these measures correspond to relatively greater polarization (i.e., a greater in-party-favoring difference in trait evaluation). In Study 1, moral polarization (*Mdn* = 13.67, IQR = 22) was larger than polarization in the domains of agency (*Mdn* = 8.83, IQR = 16.75), $p < .001$, and sociability (*Mdn* = 12.33, IQR = 17.67), $p < .001$. In Study 2, similarly, moral polarization (*Mdn* = 15, IQR = 23) was larger than polarization in both agency (*Mdn* = 9, IQR = 16.83) and sociability (*Mdn* = 10, IQR = 18.33) domains, $p < .001$ in both Wilcoxon signed-rank tests. Taken together, these results show that moral polarization was larger in magnitude than polarization manifested in the nonmoral domains.

While moral polarization is distinguishable in magnitude from polarization in the nonmoral domains, the question remains whether it reflects a distinguishable *concept*. The results of section “Nonmoral Polarization” (above) suggest that the three distinct trait indices of polarization may reflect a similar underlying factor—for example,

simple like/dislike of the target. We examined this possibility by conducting exploratory factor analyses on the trait ratings data.

First, in both study samples we conducted an exploratory factor analysis specifying a 3-factor solution with varimax rotation—one factor per trait domain—using the *psych* package in R (Revelle, 2018). We conducted separate factor analyses for trait ratings on the Democratic target and Republican target. In general, the 3-factor solution was poorly supported by the data as indicated by the factor loadings. We thus conducted a *Very Simple Structure* analysis as a guide to the optimal number of factors to extract (Revelle & Rocklin, 1979). The plots from the VSS analyses are displayed in the Appendix (Figures A.1, A.2, A.3, and A.4). They suggest that a 1-factor solution provided optimal or near-optimal fit in most cases. Therefore, in Tables A.1 and A.2 in the Appendix we report the factor loadings from exploratory factor analyses where we specified a 1-factor solution with varimax rotation. The pattern of loadings show that most of the trait ratings for each target load onto a single factor, with moral traits tending to have the highest loadings (scattered in the range of .80). However, several *non*moral trait ratings also loaded onto the factor in the range of .80 in both Studies 1 and 2 (for example, the trait "knowledgeable"). These results suggest that the three trait indices of polarization may reflect the same (or similar) underlying factor. From the pattern of signs on the factor loadings, the factor appears to be best conceived of as like/dislike of the target. We discuss the implications of this possibility in the discussion.

Beliefs About Others

Recall that, after subjects made their own choice in the IPD-MD, they reported their beliefs about what option each other player in the game—their two in-party members, and three out-party members—had chosen. Previous research suggests that patterns of ingroup favouritism are underpinned by beliefs about the differential behaviour of one's ingroup members vs. outgroup members (Brewer, 1999; Yamagishi et al., 1999). We thus preregistered our intention to investigate whether (single-index) moral polarization was associated with out-party hostility distinct from subjects' beliefs about the behaviour of the other players.

We created two new variables for this analysis. To create these variables, we first dummy-coded whether the subject believed that each in-party and out-party member expressed out-party hostility (coded 1) or not (0). We then summed these dummy variables separately for the in-party and out-party members: producing one score between 0-2, indexing the subjects' belief about the number of *in-party* members expressing out-party hostility (M_{S1} = 0.72, SD_{S1} = 0.84; M_{S2} = 0.66, SD_{S2} = 0.81); and another score between 0-3, indexing subjects' belief about the number of *out-party* members expressing out-party hostility (M_{S1} = 0.84, SD_{S1} = 1.08; M_{S2} = 0.80, SD_{S2} = 1.08). We entered these two new variables as additional predictors in the weighted-sum index models. The moral polarization ORs from these models are plotted in Figure 3 (*Beliefs about others*), and show that the association between moral polarization and out-party hostility in the IPD-MD slightly increased in size (and remained statistically significant) in both Study 1 and 2. The meta-analytic OR was 1.80, $p < .001$ [1.30, 2.50]. Interestingly, these models revealed that subjects' beliefs about the expressed out-party hostility of their two *in-party* members (but not out-party members) strongly predicted their *own* expression of out-party hostility. We return to this result below.

Further Exploratory Analyses

Our data afforded a series of exploratory analyses regarding further questions of interest. These are reported below. We note that, for each of these exploratory analyses, we exclude only those respondents with missing values on the relevant variables, as well as duplicate IDs.

Beliefs About In-Party Behaviour

As revealed in the exploratory analyses above, subjects' beliefs about the out-party hostility expressed by *in-party* members strongly predicted their *own* out-party hostility in the IPD-MD. To examine this relationship distinct from the moral polarization variable, we fitted a binomial logistic regression model where the outcome variable was out-party hostility (dummy coded as usual); and the only two predictor variables were belief about the number of (i) in-party members (0-2), and (ii) out-party members (0-3) expressing out-party hostility. As in the analysis with moral polarization, the former predictor variable was strongly associated with out-party hostility in both studies: $OR_{S1} = 5.44$, $p < .001$, 95% CI [3.67, 8.06]; $OR_{S2} = 4.81$, $p < .001$ [3.64, 6.36]. In other words, the belief that one's in-party members expressed out-party hostility shared a strong positive association with expressing out-party hostility oneself. Figure 5 displays the data upon which the models are based, and illustrates the starkness of the result. We consider this result further in the discussion. In contrast to beliefs about the in-party, subjects' beliefs about the number of *out*-party members expressing out-party hostility did not significantly predict their own expression of out-party hostility, in either study: $OR_{S1} = 1.08$, $p = .583$ [0.83, 1.40]; $OR_{S2} = 1.09$, $p = .379$ [0.90, 1.32].

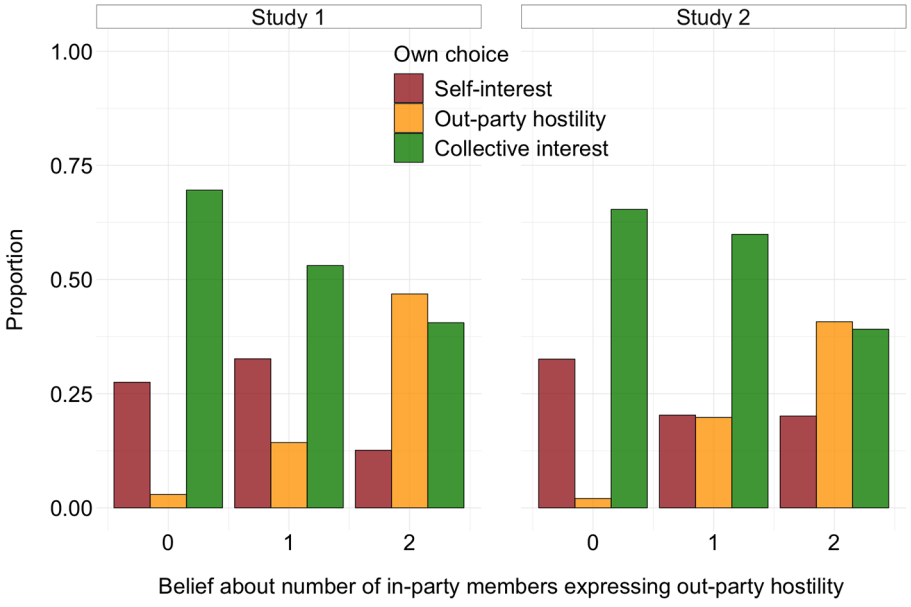

Figure 5. Proportion of choices in the IPD-MD game as a function of subjects' beliefs about the number of in-party members expressing out-party hostility.

Note. Study 1 $N = 449$ (belief 0 group $N = 240$; belief 1 group $N = 98$; belief 2 group $N = 111$); Study 2 $N = 864$ (belief 0 group $N = 488$; belief 1 group $N = 192$; belief 2 group $N = 184$).

The Ideological "Prejudice Gap"

Our data contribute to debate over the ideological "prejudice gap" (Brandt et al., 2014; Sibley & Duckitt, 2008). In particular, the ideological-conflict hypothesis (Brandt et al., 2014) predicts that people on the ideological left and ideological right exhibit approximately *symmetrical* levels of prejudice toward groups that hold values at odds with their own; as contrasted against the hypothesis of a left-right *a*symmetry in prejudicial behavior (Sibley & Duckitt, 2008). We tested these competing hypotheses by comparing rates of out-party hostility between Democratic-identifying and Republican-identifying subjects. To maximize statistical power, we pooled the data from Study 1

and Study 2 before conducting this comparison (combined N = 1,354). Among Democratic-identifying subjects, N = 111 (14.6%) expressed out-party hostility; among Republican-identifying subjects, N = 94 (15.8%) expressed out-party hostility. According to a chi-squared test, the difference was not statistically significant: $\chi^2(1) = 0.18$, p = .673. This result is inconsistent with the prejudice gap (left-right asymmetry) hypothesis.

Perceived Threat Posed by the Out-Party

We examined the association between the perception that the out-party posed a *threat* to the United States and its citizens and moral evaluation of the out-party. Recall that we collected two threat perception variables from subjects (both scored from 1-7); one concerning the "realistic" threat posed by the out-party i.e., threat to the power, safety, and resources of the US, and the other concerning "symbolic" threat; that is, threat to the values and identity of the US. The two variables were strongly correlated: $r_{S1}(451) = .78$, $p < .001$, 95% CI [.74, .81]; $r_{S2}(872) = .83$, $p < .001$ [.81, .85]. Thus, we combined them into a single threat perception variable by taking their mean (variable: perceived threat). Perceived threat was strongly negatively correlated with moral evaluation of the out-party—i.e., with the weighted-sum index of moral evaluation of the out-party target—in both studies: $r_{S1}(451) = -.56$, $p < .001$ [-.62, -.49]; $r_{S2}(868) = -.48$, $p < .001$ [-.53, -.43]. In other words, more negative beliefs about the moral character of the out-party were associated with a stronger belief that they posed a threat to the safety and values of the US and its citizens.

Discussion

We hypothesized that moral polarization would be associated with behavioural expressions of out-party hostility in the US political context. In two studies, we tested this hypothesis with large samples of US partisans and a behavioural economic game measure of outgroup hostility (Weisel & Böhm, 2015). The primary preregistered analyses were as predicted: Expressions of out-party hostility increased in conjunction with moral polarization (meta-analytic odds ratio = 1.59, 95% CI [1.19, 2.12]). In a series of subsequent preregistered and exploratory sensitivity analyses, we tested the robustness of this result. While these analyses indicated that the primary preregistered result was somewhat robust, they also highlighted important constraints on the inference that moral polarization is associated with out-party hostility in the US political context. We consider the implications of these and our various other results below.

We observed important exceptions to the general robustness of our primary preregistered result. Most notably, the *interaction* between in-party moral evaluation and out-party moral evaluation did not corroborate the model results in which moral polarization was construed as a single index. We discovered that this mismatch was due to the fact that the single-index of moral polarization seemed to predict out-party hostility via main effects of in-party moral evaluation (Study 1) and out-party moral evaluation (Study 2) *per se*. In other words, the perceived moral "gap" between parties was not as important as the moral evaluation of one party or the other. We take this result to be contrary to our hypothesis, which was that partisans who *both* (a) morally championed the in-party *and* (b) morally demonized the out-party would be most likely to express out-party hostility. Given the inconsistency of this result across studies, and the fact that similar past work on affective polarization used a single-index measure (Iyengar et al., 2012; Lelkes & Westwood, 2017), more work is necessary to determine with confidence whether in- or out-party evaluation is decisive in explaining variance in behavioural expressions of out-party hos-

tility. Nevertheless, our results with the single-index of moral polarization corroborate previous work on affective polarization and partisan prejudice—work to which we now turn.

Lelkes and Westwood (2017) find evidence that even those partisans who are the most affectively polarized are generally unwilling to endorse discriminatory behaviour against the political opposition (see also Westwood, Peterson, & Lelkes, 2018). Our results extend their findings in two ways. First, we find that the same pattern holds when using an incentivized, *behavioural* measure of out-party hostility, rather than self-report (as those authors used). More specifically, our primary preregistered results indicated that even partisans at the upper limit of moral polarization were relatively unlikely to exhibit out-party hostility (predicted probabilities of 0.21 and 0.17 in Studies 1 and 2, respectively). Furthermore, as indicated by these predicted probabilities, levels of out-party hostility were low in absolute terms as well (see Table 2)[iv]. Second, we measured *moral* polarization, rather than generalized affective polarization. Given that we identified moral polarization to be greater in magnitude than polarization in nonmoral domains of evaluation—if not conceptually distinct (see below)—it is possible that our studies provided fertile conditions for a stronger association between affective polarization and partisan prejudice (out-party hostility) to emerge[v]. Taking our results together with the large sample sizes in our studies, and the rather *mild* form of out-party hostility afforded by the IPD-MD, suggests that the association between affective polarization and out-party hostility in the US political context is small and somewhat tenuous (cf. Lelkes & Westwood, 2017; Westwood, Peterson, & Lelkes, 2018).

Notwithstanding this convergence in findings, however, there is a particular limitation of our studies that warrants mention and precludes a strong interpretation of our results along the foregoing lines. That is, our sampling population. We recruited subjects from Amazon's Mechanical Turk, a survey platform whose subjects are known to fall short of demographically-representing the wider US population (Chandler & Shapiro, 2016). As highlighted in the introduction, Mason (2016, 2018) finds that US party identity is increasingly in alignment with demographic identities (e.g., race, religiosity), and, importantly, that this alignment may serve to weaken barriers to out-party hostility (Mason & Wronski, 2018; Roccas & Brewer, 2002). For this reason, insofar as our subjects did not faithfully represent the demographic identities of the wider US population, it is possible that our analyses mis-estimated the population-level association between moral polarization and out-party hostility in IPD-MD. Ultimately, though, we consider this minimally problematic for our overall interpretation of our results, given that (i) there is no evidence that a more faithful demographic representation would have strengthened the target association—it may just as well have *attenuated* it—and (ii) the results of Lelkes and Westwood (2017), that converge with our own, *are* based on representative samples of US adults.

In contrast to the equivocal association between moral polarization and out-party hostility, we observed relatively stronger evidence of moral polarization *per se*. Specifically, on the preregistered correlational index (Table 3 and Figure 2), and exploratory weighted-sum (Figure 2) and trait-summary indices (Figure 4), moral polarization among US partisans appeared large and robust. Furthermore, we found evidence that moral polarization was greater in magnitude than polarization observed in the *non*moral domains of evaluation. Despite this, we are unable to conclude that moral polarization is conceptually distinct from nonmoral polarization; exploratory factor analyses suggested that the three trait indices of polarization reflect the same (or similar) underlying factor. From the pattern of signs on the factor loadings, the factor appears to be best conceived of as like/dislike of the target. These results help explain why moral polarization did not predict out-party hostility distinct from polarization in the nonmoral domains—because they suggest that moral polarization is not a distinct concept. Instead, it may simply best reflect

an underlying factor of partisan like/dislike. We consider two interpretations of this result as it relates to the phenomenon of affective polarization (Iyengar et al., 2012; Iyengar et al., 2018).

One interpretation is that moral polarization is simply a more proximate indicator of whatever underlying construct is manifesting as affective polarization. For example, assuming that domain-general partisan like/dislike is the underlying construct, one would expect moral polarization to be stronger than nonmoral polarization for the reason that moral traits share a stronger relationship with liking and respecting other people/groups than do nonmoral traits (Hartley et al., 2016). In other words, partisans express their dislike of the out-party and liking of the in-party through whichever route is available; and moral (vs. nonmoral) evaluation just happens to be more "cathartic" in this sense.

On the other hand, it seems likely that moral evaluation also *causes* the (dis-)liking of other people/groups. This is implied by a long programme of research showing that moral content guides—and, in fact, dominates—humans' global evaluations of other people and groups, ostensibly because the moral character/benevolence of others can have a very direct and consequential impact on one's own wellbeing (reviewed in Wojciszke, 2005). On this view, affective polarization *in general* may be a function of moral polarization *in particular*. That is, partisans "like" the in-party and "dislike" the out-party in part *because* the former are perceived to be fairer, more trustworthy, less prejudiced—in other words, more *benevolent*—than the latter. This perspective accords well with the distinct role of moral psychology in contemporary American politics, as outlined in the introduction (Brady et al., 2017; Koleva et al., 2012; Ryan, 2014, 2017), as well as the apparent moderating effect of moral conviction on affective polarization (Garrett & Bankert, 2018).

Unfortunately, which of the foregoing interpretations is ultimately correct cannot be determined on the basis of the current data. Future work might adjudicate by experimentally assigning moral and nonmoral characteristics to in- and out-party targets, and measuring affective polarization. Nevertheless, our results *do* indicate that estimates of the magnitude of affective polarization inferred via trait measures (e.g., Iyengar et al., 2012; Levendusky, 2018) depend non-trivially on the type of traits used (see also Levendusky, 2018). Previous work that mixes moral and nonmoral traits may thus have underestimated affective polarization.

Our findings regarding trait polarization also relate to work on political dehumanization. For example, Crawford, Modri, and Motyl (2013) used target trait ascriptions to infer dehumanization of one's political opponents among US partisans. These authors found some evidence that out-party antipathy accounted for variance in dehumanization of the out-party on these trait ratings. Does this suggest that moral trait polarization (our measure) is associated with out-party hostility via political dehumanization? Perhaps. However, recent work has drawn a distinction between dehumanization on the one hand, and morality-based aggression on the other. Specifically, while denying the humanity of others may allow people to "look the other way" when intergroup aggression is committed for *material* benefit—like in the lucrative enterprise of slavery (Bruneau & Kteily, 2017)—dehumanization seems less well equipped to explain *moral* aggression, because it would seem to rob outgroup members of moral responsibility and thus moral condemnation (Rai et al., 2017). Indeed, nonhuman animals, robots, and objects are seldom the subjects of moral blame. While some work indicates that people dehumanize those deserving of punishment (Khamitov et al., 2016), recent evidence suggests that morally motivated perpetrators may also *humanize* others to justify aggression against them (Rai et al., 2017); a proposition consistent with earlier research showing that dehumanization renders people *less* susceptible to moral blame (Bastian et al., 2011). More—ideally experimen-

tal—work is necessary to understand how political dehumanization relates to moral (or affective) polarization and behavioural expressions of out-party hostility.

We observed a strong positive association between subjects' beliefs about the number of *in-party* members expressing out-party hostility and their *own* expression of out-party hostility (Figure 5). Though this association was observed in exploratory analyses—and must be interpreted as such—we note that the relevant odds ratios in both studies survive Bonferroni corrections of 1×10^{14} to the *p*-values; implying that the association is robust. We offer two explanations for this intriguing result. The first and we think more likely explanation is that subjects *projected* their own behaviour in the IPD-MD onto their judgment of what the in-party members would do. A long line of research demonstrates that people engage in "social projection" of this kind when asked to make information-deprived judgments of other people (reviewed in Krueger, 2007; Robbins & Krueger, 2005). The logic behind the utility of social projection is that—because most people are in the majority most of the time—projection allows people to make quick and reasonably accurate judgments (on average) about unknown others (Krueger, 2007; Krueger & Chen, 2014). Indeed, in our studies subjects received only sparse information about the other players (i.e., only their party affiliation); providing good conditions for social projection.

An alternative explanation for the result is that subjects tailored their own out-party hostility behaviour to what they believed the other players in the IPD-MD would do. Specifically, to whether they believed the *in-party* would express out-party hostility; akin to a reciprocation- or conformity-type effect. We think this explanation is less likely than social projection. Primarily because a large body of evidence shows that projection to *in*group members is typically greater than projection to *out*group members (for a meta-analysis, see Robbins & Krueger, 2005). This is strongly consistent with our results, where subjects' beliefs about the out-party hostility expressed by *out-party members* were only trivially associated with their own out-party hostility behaviour. The alternative explanation—the notion that subjects tailored their behaviour to the expected behaviour of the other players—appears less able to explain this non-association. This is because, assuming this alternative explanation is right, one would expect that beliefs about the expressed out-party hostility of the *out-party members* would, to some extent at least, also affect subjects' own choice to express out-party hostility. For example, it is reasonable to expect that they would be positively correlated—reflecting a desire for "pre-emptive strike" (Böhm et al., 2016; Simunovic et al., 2013); in other words, if I think the out-party will aggress against me, I am more inclined to aggress against them. That we did not observe such an association provides some evidence that subjects were not tailoring their own out-party hostility behaviour to what they believed the other players in the IPD-MD would do.

Regardless of which explanation is actually right, the result itself highlights a potentially fruitful avenue by which to predict—ahead of time and with reasonable accuracy—the out-party hostility behaviour of partisans. Namely, to query whether they believe that the typical in-party member would express out-party hostility. This may be a particularly useful strategy to identify those most likely to express out-party hostility where there exist disincentives to answering in the affirmative oneself. We leave it to future research to explore this idea.

In this paper, we investigated the association between moral polarization—the tendency for people to view opposing partisans' moral character negatively, and co-partisans' moral character positively—and behavioural expressions of out-party hostility in the US political context. Our results strike an optimistic chord: Taken together, they suggest that the association is probably small and somewhat tenuous. Though moral polarization itself appeared large—and may exceed prior estimates of trait affective polarization—in our sample even the *most* morally polarized partisans were reluctant to engage in a rather mild form of hostile behaviour toward the out-party. These findings converge

with recent evidence that polarization—moral or otherwise—has yet to translate into the average US partisan wanting to express hostile and directly discriminatory behaviour toward their out-party counterparts.

Notes

i) In the preregistered protocols, we referred to this decision option as "parochial altruism". However, here we refer to it as "outgroup (out-party) hostility" to clearly distinguish between *our* focus—which is simply those instances where ingroup "love" and outgroup hostility appear in conjunction (such as in suicide terrorism and war)—and the parochial altruism *hypothesis*—which concerns the evolutionary origins of this conjunction. While the latter hypothesis has received recent criticism (e.g., Rusch et al., 2016; Yamagishi & Mifune, 2016), these criticisms do not contest the *existence* of ingroup love/outgroup hostility, but, rather, the proposition that the conjunction of these behaviours manifests (i) consistently at the individual-level (i.e., as a within-individual correlation) and (ii) as a result of group-level selection pressure (for more detailed discussion, we refer to Rusch et al., 2016; Yamagishi & Mifune, 2016). We are grateful to an anonymous reviewer for emphasizing this point.

ii) We are grateful to Mark Brandt and two anonymous reviewers for pointing out where and why alternative measures of moral polarization might provide for more valid inferences—and to Reviewer #2 in particular for suggesting the weighted-sum index of moral polarization (see in-text).

iii) In the preregistered protocols, we referred to this variable as "inframoralization". However, here we changed the label to "moral polarization" for descriptive clarity and consistency with concepts as defined in closely relevant work (Iyengar et al., 2012; Iyengar & Westwood, 2015). The variable is unchanged in all other respects. We thank Mark Brandt for emphasizing the relevance of this work to the present investigation.

iv) Of course, social desirability bias may to some extent account for these low numbers; a concern we are unable to quantify and/or rule out here. Although we note Lelkes and Westwood's (2017) point that—unlike prejudice based on race or other such characteristics—political prejudice is less constrained by social desirability bias. We recognize this and thus expect that while there may be some residual bias, it would not be enough to drastically change our results. A further factor that may possibly help explain the low rates of out-party hostility is that the study procedure asked subjects to provide trait ratings prior to deciding in the IPD-MD. If subjects were fatigued by the decision point, they may not have paid as much attention to their decision and/or been more likely to choose the self-interested option (to maximize earnings).

v) However, we acknowledge that feeling thermometer ratings are perhaps more "emotional" than trait ratings, and thus the former may be more likely to predict expressions of out-party hostility *per se*. To our knowledge, though, there has yet to be a systematic comparison between (a) trait measures and (b) feeling-thermometer measures of affective polarization in predicting behavioural prejudice.

Funding

We are grateful to the Economic and Social Research Council for funding this project (grant no. ES/J500148/1).

Competing Interests

The authors have declared that no competing interests exist.

Acknowledgments

The authors have no support to report.

Data Availability

For this paper a dataset and preregistered protocols for both studies are freely available (see the Supplementary Materials section).

Supplementary Materials

The raw data and analysis scripts to reproduce the results and figures reported in this paper are available online via the project hub on the Open Science Framework: https://osf.io/mceqh/.

Both studies were preregistered on *AsPredicted*: https://aspredicted.org/e3hw9.pdf (link to Study 1 protocol); https://aspredicted.org/tiuw7.pdf (Study 2 protocol).

Index of Supplementary Materials

Tappin, B. M., & McKay, R. T. (2019). *Moral polarization and out-party hostility in the US political context* [Supplementary materials]. https://osf.io/mceqh/

McKay, R. T., & Tappin, B. (2016). *Inframoralization predicts parochial altruism. (#995)* [Preregistered study protocol]. https://aspredicted.org/e3hw9.pdf

Tappin, B., & McKay, R. T. (2016). *Inframoralization predicts parochial altruism, replication (#1123)* [Preregistered study protocol]. https://aspredicted.org/tiuw7.pdf

References

Arechar, A. A., Gächter, S., & Molleman, L. (2018). Conducting interactive experiments online. *Experimental Economics, 21*, 99-131. https://doi.org/10.1007/s10683-017-9527-2

Arnold, J. B. (2017). ggthemes: Extra themes, scales and geoms for 'ggplot2' (R package version 3.4.0) [Computer software]. Retrieved from https://CRAN.R-project.org/package=ggthemes

Auguie, B. (2017). gridExtra: Miscellaneous functions for "grid" graphics (R package version 2.3) [Computer software]. Retrieved from https://CRAN.R-project.org/package=gridExtra

Bastian, B., Laham, S. M., Wilson, S., Haslam, N., & Koval, P. (2011). Blaming, praising, and protecting our humanity: The implications of everyday dehumanization for judgments of moral status. *British Journal of Social Psychology, 50*, 469-483. https://doi.org/10.1348/014466610X521383

Bilewicz, M., & Vollhardt, J. R. (2012). Evil transformations: Social-psychological processes underlying genocide and mass killing. In A. Golec de Zavala & A. Cichocka (Eds.), *Social psychology of social problems: The intergroup context* (pp. 280-307). New York, NY, USA: Palgrave Macmillan.

Böhm, R., Rusch, H., & Gürerk, Ö. (2016). What makes people go to war? Defensive intentions motivate retaliatory and preemptive intergroup aggression. *Evolution and Human Behavior, 37*, 29-34. https://doi.org/10.1016/j.evolhumbehav.2015.06.005

Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., & Van Bavel, J. J. (2017). Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences of the United States of America, 114*, 7313-7318. https://doi.org/10.1073/pnas.1618923114

Brandt, M. J. (2017). Predicting ideological prejudice. *Psychological Science, 28*, 713-722. https://doi.org/10.1177/0956797617693004

Brandt, M. J., Reyna, C., Chambers, J. R., Crawford, J. T., & Wetherell, G. (2014). The ideological-conflict hypothesis: Intolerance among both liberals and conservatives. *Current Directions in Psychological Science, 23*, 27-34. https://doi.org/10.1177/0963721413510932

Brewer, M. B. (1999). The psychology of prejudice: Ingroup love and outgroup hate? *Journal of Social Issues, 55*, 429-444. https://doi.org/10.1111/0022-4537.00126

Bruneau, E., & Kteily, N. (2017). The enemy as animal: Symmetric dehumanization during asymmetric warfare. *PLoS One, 12*, Article e0181422. https://doi.org/10.1371/journal.pone.0181422

Carlin, R. E., & Love, G. J. (2013). The politics of interpersonal trust and reciprocity: An experimental approach. *Political Behavior, 35*, 43-63. https://doi.org/10.1007/s11109-011-9181-x

Carlin, R. E., & Love, G. J. (2018). Political competition, partisanship and interpersonal trust in electoral democracies. *British Journal of Political Science, 48*, 115-139. https://doi.org/10.1017/S0007123415000526

Chandler, J., & Shapiro, D. (2016). Conducting clinical research using crowdsourced convenience samples. *Annual Review of Clinical Psychology, 12*, 53-81. https://doi.org/10.1146/annurev-clinpsy-021815-093623

Clifford, S., Jewell, R. M., & Waggoner, P. D. (2015). Are samples drawn from Mechanical Turk valid for research on political ideology? *Research & Politics, 2*, Article 2053168015622072. https://doi.org/10.1177/2053168015622072

Crawford, J. T., Brandt, M. J., Inbar, Y., Chambers, J. R., & Motyl, M. (2017). Social and economic ideologies differentially predict prejudice across the political spectrum, but social issues are most divisive. *Journal of Personality and Social Psychology, 112*, 383-412. https://doi.org/10.1037/pspa0000074

Crawford, J., Modri, S., & Motyl, M. (2013). Bleeding-heart liberals and hard-hearted conservatives: Subtle political dehumanization through differential attributions of human nature and human uniqueness traits. *Journal of Social and Political Psychology, 1*, 86-104. https://doi.org/10.5964/jspp.v1i1.184

Dowle, M., & Srinivasan, A. (2017). Data.table: Extension of 'data. frame' (R package version 1.10.4-3) [Computer software]. Retrieved from https://CRAN.R-project.org/package=data.table

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*, 1149-1160. https://doi.org/10.3758/BRM.41.4.1149

Garrett, K. N., & Bankert, A. (2018). The moral roots of partisan division: How moral conviction heightens affective polarization. *British Journal of Political Science*. Advance online publication. https://doi.org/10.1017/S000712341700059X

Giner-Sorolla, R., Leidner, B., & Castano, E. (2012). Dehumanization, demonization, and morality shifting: Paths to moral certainty in extremist violence. In M. A. Hogg & D. L. Blaylock (Eds.), *Extremism and the psychology of uncertainty* (pp. 165-182). Chichester, United Kingdom: Wiley-Blackwell.

Ginges, J., Atran, S., Sachdeva, S., & Medin, D. (2011). Psychology out of the laboratory: The challenge of violent extremism. *The American Psychologist, 66*, 507-519. https://doi.org/10.1037/a0024715

Goodwin, G. P., Piazza, J., & Rozin, P. (2014). Moral character predominates in person perception and evaluation. *Journal of Personality and Social Psychology, 106*, 148-168. https://doi.org/10.1037/a0034726

Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology, 96*(5), 1029-1046. https://doi.org/10.1037/a0015141

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

Halevy, N., Bornstein, G., & Sagiv, L. (2008). “In-group love” and “outgroup hate” as motives for individual participation in intergroup conflict: A new game paradigm. *Psychological Science, 19*, 405-411. https://doi.org/10.1111/j.1467-9280.2008.02100.x

Halperin, E. (2008). Group-based hatred in intractable conflict in Israel. *Journal of Conflict Resolution, 52*, 713-736. https://doi.org/10.1177/0022002708314665

Hartley, A. G., Furr, R. M., Helzer, E. G., Jayawickreme, E., Velasquez, K. R., & Fleeson, W. (2016). Morality’s centrality to liking, respecting, and understanding others. *Social Psychological & Personality Science, 7*, 648-657. https://doi.org/10.1177/1948550616655359

Hothorn, T., Hornik, K., van de Wiel, M. A., & Zeileis, A. (2008). Implementing a class of permutation tests: The coin package. *Journal of Statistical Software, 28*, 1-23. Retrieved from http://www.jstatsoft.org/v28/i08/https://doi.org/10.18637/jss.v028.i08

Huber, G. A., & Malhotra, N. (2017). Political homophily in social relationships: Evidence from online dating behavior. *The Journal of Politics, 79*, 269-283. https://doi.org/10.1086/687533

Iyengar, S., Lelkes, Y., Levendusky, M., Malhotra, N., & Westwood, S. (2018). The origins and consequences of affective polarization. *Annual Review of Political Science, 22*. Advance online publication. https://doi.org/10.1146/annurev-polisci-051117-073034

Iyengar, S., Sood, G., & Lelkes, Y. (2012). Affect, not ideology: Social identity perspective on polarization. *Public Opinion Quarterly, 76*, 405-431. https://doi.org/10.1093/poq/nfs038

Iyengar, S., & Westwood, S. J. (2015). Fear and loathing across party lines: New evidence on group polarization. *American Journal of Political Science, 59*, 690-707. https://doi.org/10.1111/ajps.12152

Jost, J. T., Federico, C. M., & Napier, J. L. (2009). Political ideology: Its structure, functions, and elective affinities. *Annual Review of Psychology, 60*, 307-337. https://doi.org/10.1146/annurev.psych.60.110707.163600

Khamitov, M., Rotman, J. D., & Piazza, J. (2016). Perceiving the agency of harmful agents: A test of dehumanization versus moral typecasting accounts. *Cognition, 146*, 33-47. https://doi.org/10.1016/j.cognition.2015.09.009

Kinder, D. R., & Kalmoe, N. P. (2017). *Neither liberal nor conservative: Ideological innocence in the American public*. Chicago, IL, USA: University of Chicago Press.

Koleva, S. P., Graham, J., Iyer, R., Ditto, P. H., & Haidt, J. (2012). Tracing the threads: How five moral concerns (especially Purity) help explain culture war attitudes. *Journal of Research in Personality, 46*, 184-194. https://doi.org/10.1016/j.jrp.2012.01.006

Koonz, C. (2003). *The Nazi conscience*. Cambridge, MA, USA: Harvard University Press.

Krueger, J. I. (2007). From social projection to social behaviour. *European Review of Social Psychology, 18*, 1-35. https://doi.org/10.1080/10463280701284645

Krueger, J. I., & Chen, L. J. (2014). The first cut is the deepest: Effects of social projection and dialectical bootstrapping on judgmental accuracy. *Social Cognition, 32*, 315-336. https://doi.org/10.1521/soco.2014.32.4.315

Leach, C. W., Ellemers, N., & Barreto, M. (2007). Group virtue: The importance of morality (vs. competence and sociability) in the positive evaluation of in-groups. *Journal of Personality and Social Psychology, 93*, 234-249. https://doi.org/10.1037/0022-3514.93.2.234

Lelkes, Y., & Westwood, S. J. (2017). The limits of partisan prejudice. *The Journal of Politics, 79*, 485-501. https://doi.org/10.1086/688223

Levendusky, M. S. (2018). Americans, not partisans: Can priming American national identity reduce affective polarization? *The Journal of Politics, 80*, 59-70. https://doi.org/10.1086/693987

Lewis, G. J., Kandler, C., & Riemann, R. (2014). Distinct heritable influences underpin in-group love and out-group derogation. *Social Psychological & Personality Science, 5*, 407-413. https://doi.org/10.1177/1948550613504967

Mason, L. (2016). A cross-cutting calm: How social sorting drives affective polarization. *Public Opinion Quarterly, 80*, 351-377. https://doi.org/10.1093/poq/nfw001

Mason, L. (2018). *Uncivil agreement: How politics became our identity*. Chicago, IL, USA: University of Chicago Press.

Mason, L., & Wronski, J. (2018). One tribe to bind them all: How our social group attachments strengthen partisanship. *Political Psychology, 39*(S1), 257-277. https://doi.org/10.1111/pops.12485

McConnell, C., Margalit, Y., Malhotra, N., & Levendusky, M. (2018). The economic consequences of partisanship in a polarized era. *American Journal of Political Science, 62*, 5-18. https://doi.org/10.1111/ajps.12330

Parker, M. T., & Janoff-Bulman, R. (2013). Lessons from morality-based social identity: The power of outgroup "hate," not just ingroup "love". *Social Justice Research, 26*, 81-96. https://doi.org/10.1007/s11211-012-0175-6

Rai, T. S., Valdesolo, P., & Graham, J. (2017). Dehumanization increases instrumental violence, but not moral violence. *Proceedings of the National Academy of Sciences of the United States of America, 114*, 8511-8516. https://doi.org/10.1073/pnas.1705238114

Rand, D. G. (2012). The promise of Mechanical Turk: How online labor markets can help theorists run behavioral experiments. *Journal of Theoretical Biology, 299*, 172-179. https://doi.org/10.1016/j.jtbi.2011.03.004

R Core Team. (2017). R: A language and environment for statistical computing [Computer software]. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from https://www.R-project.org/

Reicher, S., Haslam, S. A., & Rath, R. (2008). Making a virtue of evil: A five-step social identity model of the development of collective hate. *Social and Personality Psychology Compass, 2*, 1313-1344. https://doi.org/10.1111/j.1751-9004.2008.00113.x

Revelle, W. (2018). Psych: Procedures for psychological, psychometric, and personality research [Computer software]. Retrieved from https://CRAN.R-project.org/package=psych

Revelle, W., & Rocklin, T. (1979). Very simple structure: An alternative procedure for estimating the optimal number of interpretable factors. *Multivariate Behavioral Research, 14*, 403-414. https://doi.org/10.1207/s15327906mbr1404_2

Robbins, J. M., & Krueger, J. I. (2005). Social projection to ingroups and outgroups: A review and meta-analysis. *Personality and Social Psychology Review, 9*, 32-47. https://doi.org/10.1207/s15327957pspr0901_3

Roccas, S., & Brewer, M. B. (2002). Social identity complexity. *Personality and Social Psychology Review, 6*, 88-106. https://doi.org/10.1207/S15327957PSPR0602_01

RStudio Team. (2016). RStudio: Integrated development for R [Computer software]. Boston, MA, USA: RStudio, Inc. Retrieved from http://www.rstudio.com

Rusch, H., Böhm, R., & Herrmann, B. (2016). Parochial altruism: Pitfalls and prospects. *Frontiers in Psychology, 7*, Article 1004. https://doi.org/10.3389/fpsyg.2016.01004

Ryan, T. J. (2014). Reconsidering moral issues in politics. *The Journal of Politics, 76*, 380-397. https://doi.org/10.1017/S0022381613001357

Ryan, T. J. (2017). No compromise: Political consequences of moralized attitudes. *American Journal of Political Science, 61*, 409-423. https://doi.org/10.1111/ajps.12248

Saucier, G., Akers, L. G., Shen-Miller, S., Knežević, G., & Stankov, L. (2009). Patterns of thinking in militant extremism. *Perspectives on Psychological Science, 4*, 256-271. https://doi.org/10.1111/j.1745-6924.2009.01123.x

Sibley, C. G., & Duckitt, J. (2008). Personality and prejudice: A meta-analysis and theoretical review. *Personality and Social Psychology Review, 12*, 248-279. https://doi.org/10.1177/1088868308319226

Simunovic, D., Mifune, N., & Yamagishi, T. (2013). Preemptive strike: An experimental study of fear-based aggression. *Journal of Experimental Social Psychology, 49*, 1120-1123. https://doi.org/10.1016/j.jesp.2013.08.003

Skitka, L. J., Bauman, C. W., & Sargis, E. G. (2005). Moral conviction: Another contributor to attitude strength or something more? *Journal of Personality and Social Psychology, 88*, 895-917. https://doi.org/10.1037/0022-3514.88.6.895

Skitka, L. J., & Mullen, E. (2002). The dark side of moral conviction. *Analyses of Social Issues and Public Policy, 2*, 35-41. https://doi.org/10.1111/j.1530-2415.2002.00024.x

Stagnaro, M. N., Dunham, Y., & Rand, D. G. (2018). Profit versus prejudice: harnessing self-interest to reduce in-group bias. *Social Psychological & Personality Science, 9*, 50-58. https://doi.org/10.1177/1948550617699254

Stephan, W. G., Ybarra, O., & Morrison, K. R. (2011). Intergroup threat theory. In T. Nelson (Ed.), *Handbook of prejudice* (pp. 43-55). New York, NY, USA: Taylor & Francis Group.

Tajfel, H., Billig, M. G., Bundy, R. P., & Flament, C. (1971). Social categorization and intergroup behaviour. *European Journal of Social Psychology, 1*, 149-178. https://doi.org/10.1002/ejsp.2420010202

Tappin, B. M., & McKay, R. T. (2017). The illusion of moral superiority. *Social Psychological & Personality Science, 8*, 623-631. https://doi.org/10.1177/1948550616673878

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36*, 1-48. Retrieved from http://www.jstatsoft.org/v36/i03/https://doi.org/10.18637/jss.v036.i03

Weisel, O. (2015). Negative and positive externalities in intergroup conflict: Exposure to the opportunity to help the outgroup reduces the inclination to harm it. *Frontiers in Psychology, 6*, Article 1594. https://doi.org/10.3389/fpsyg.2015.01594

Weisel, O., & Böhm, R. (2015). “Ingroup love” and “outgroup hate” in intergroup conflict between natural groups. *Journal of Experimental Social Psychology, 60*, 110-120. https://doi.org/10.1016/j.jesp.2015.04.008

Westwood, S. J., Iyengar, S., Walgrave, S., Leonisio, R., Miller, L., & Strijbis, O. (2018). The tie that divides: Cross-national evidence of the primacy of partyism. *European Journal of Political Research, 57*, 333-354. https://doi.org/10.1111/1475-6765.12228

Westwood, S. J., Peterson, E., & Lelkes, Y. (2018). *Are there still limits on partisan prejudice?* (Working paper). Retrieved from https://www.dartmouth.edu/~seanjwestwood/papers/stillLimits.pdf

Wickham, H. (2007). Reshaping data with the reshape package. *Journal of Statistical Software, 21*. https://doi.org/10.18637/jss.v021.i12

Wickham, H. (2011). The split-apply-combine strategy for data analysis. *Journal of Statistical Software, 40*. https://doi.org/10.18637/jss.v040.i01

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. New York, NY, USA: Springer.

Wickham, H. (2018). Scales: Scale functions for visualization (R package version 1.0.0) [Computer software]. Retrieved from https://CRAN.R-project.org/package=scales

Wickham, H., François, R., Henry, L., & Müller, K. (2018). Dplyr: A grammar of data manipulation (R package version 0.7.6) [Computer software]. Retrieved from https://CRAN.R-project.org/package=dplyr

Wojciszke, B. (2005). Morality and competence in person-and self-perception. *European Review of Social Psychology, 16*, 155-188. https://doi.org/10.1080/10463280500229619

Yamagishi, T., Jin, N., & Kiyonari, T. (1999). Bounded generalized reciprocity. *Advances in Group Processes, 16*, 161-197.

Yamagishi, T., & Mifune, N. (2016). Parochial altruism: Does it explain modern human group psychology? *Current Opinion in Psychology, 7*, 39-43. https://doi.org/10.1016/j.copsyc.2015.07.015

Appendix

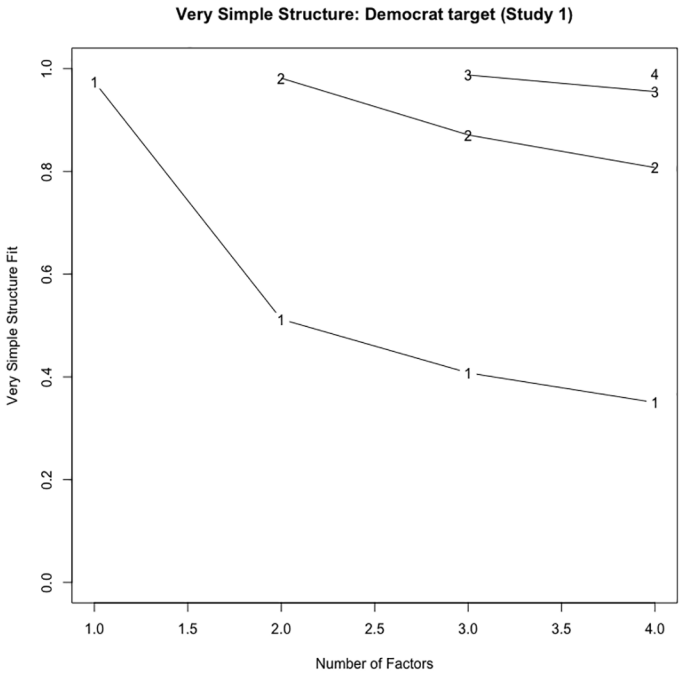

Figure A.1. Very simple structure analysis of Democrat target trait ratings in Study 1.

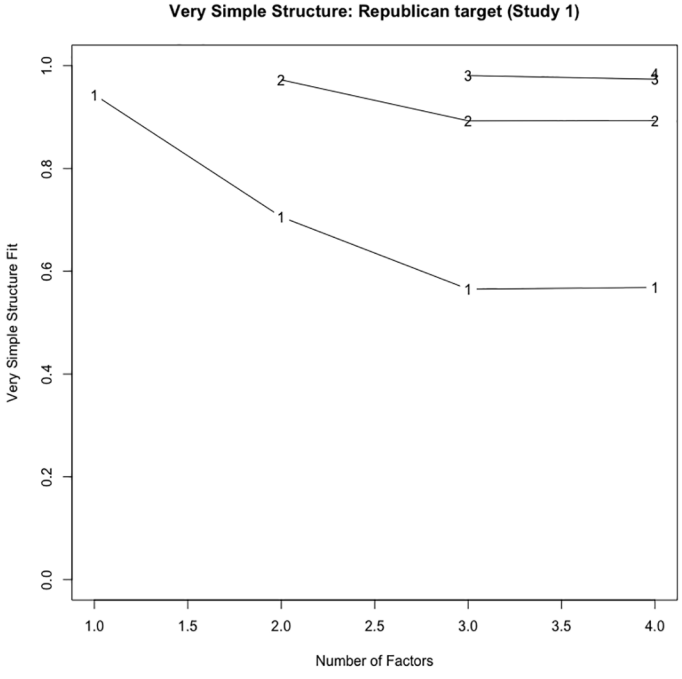

Figure A.2. Very simple structure analysis of Republican target trait ratings in Study 1.

Table A.1

Factor Loadings From Exploratory Factor Analysis (Study 1)

Trait	Democrat target	Republican target
Hardworking	-0.79	0.64
Knowledgeable	-0.81	0.76
Competent	-0.82	0.77
Creative	-0.63	0.68
Determined	-0.52	0.34
Lazy	0.76	-0.52
Undedicated	0.70	-0.47
Unintelligent	0.78	-0.73
Unmotivated	0.68	-0.40
Illogical	0.81	-0.78
Sociable	-0.58	0.60
Cooperative	-0.84	0.79
Warm	-0.83	0.83
Family-orientated	-0.73	0.52
Easygoing	-0.70	0.72
Cold	0.81	-0.76
Disagreeable	0.79	-0.78
Rude	0.83	-0.79
Humorless	0.69	-0.75
Uptight	0.70	-0.66
Honest	-0.83	0.81
Trustworthy	-0.82	0.85
Fair	-0.84	0.87
Respectful	-0.86	0.85
Principled	-0.79	0.65
Insincere	0.82	-0.81
Prejudiced	0.74	-0.80
Disloyal	0.79	-0.59
Manipulative	0.81	-0.80
Deceptive	0.82	-0.81

Note. Each column corresponds to a factor analysis on the respective target ratings with the number of factors to extract set to 1. Rotation is set to *varimax*.

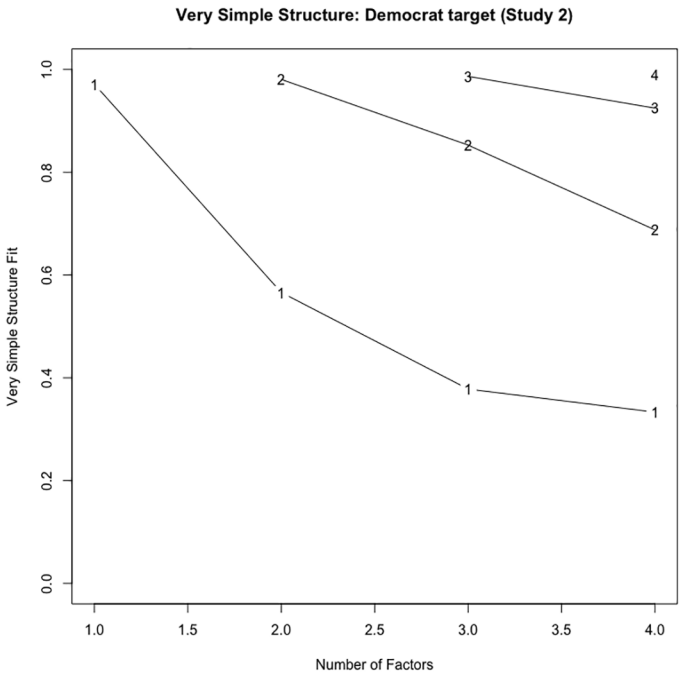

Figure A.3. Very simple structure analysis of Democrat target trait ratings in Study 2.

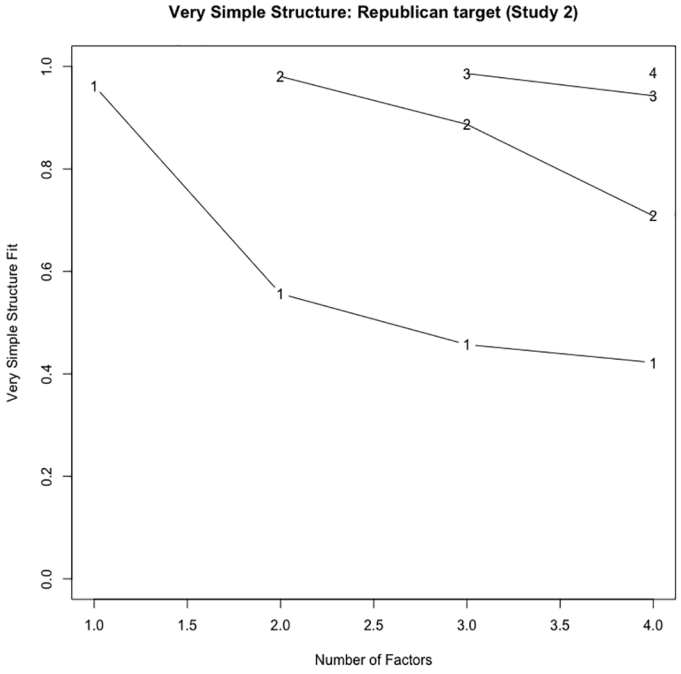

Figure A.4. Very simple structure analysis of Republican target trait ratings in Study 2.

Table A.2

Factor Loadings From Exploratory Factor Analysis (Study 2)

Trait	Democrat target	Republican target
Knowledgeable	-0.81	-0.81
Determined	-0.58	-0.42
Intelligent	-0.80	-0.81
Creative	-0.69	-0.75
Organized	-0.66	-0.64
Incompetent	0.82	0.83
Lazy	0.79	0.66
Unmotivated	0.75	0.57
Unproductive	0.83	0.74
Weak	0.77	0.67
Sociable	-0.64	-0.69
Easygoing	-0.68	-0.74
Playful	-0.56	-0.62
Happy	-0.72	-0.73
Funny	-0.60	-0.62
Disagreeable	0.79	0.80
Negative	0.83	0.86
Reckless	0.79	0.77
Humorless	0.73	0.77
Uptight	0.70	0.73
Honest	-0.84	-0.84
Trustworthy	-0.86	-0.88
Just	-0.85	-0.83
Fair	-0.84	-0.86
Principled	-0.80	-0.72
Violent	0.71	0.70
Insincere	0.82	0.84
Greedy	0.81	0.79
Prejudiced	0.72	0.82
Disloyal	0.79	0.69

Note. Each column corresponds to a factor analysis on the respective target ratings with the number of factors to extract set to 1. Rotation is set to *varimax*.