PHILOSOPHICAL TRANSACTIONS B

royalsocietypublishing.org/journal/rstb

Research

Cite this article: Lees J, Cikara M. 2021 Understanding and combating misperceived polarization. *Phil. Trans. R. Soc. B* **376**: 20200143.
https://doi.org/10.1098/rstb.2020.0143

Accepted: 1 October 2020

One contribution of 18 to a theme issue 'The political brain: neurocognitive and computational mechanisms'.

Subject Areas:
cognition

Keywords:
polarization, intergroup relations, meta-perception, social psychology, politics

Author for correspondence:
Jeffrey Lees
e-mail: jeffrey.m.lees@gmail.com

THE ROYAL SOCIETY PUBLISHING

Understanding and combating misperceived polarization

Jeffrey Lees[1] and Mina Cikara[2]

[1]Department of Economics, Clemson University, Wilbur O. and Ann Powers Hall, Clemson, SC 29634, USA
[2]Department of Psychology, Harvard University, William James Hall, Cambridge, MA 02138, USA

JL, 0000-0001-6030-4207; MC, 0000-0002-6612-4474

By many accounts politics is becoming more polarized, yielding dire consequences for democracy and trust in government. Yet a growing body of research on so-called false polarization finds that perceptions of 'what the other side believes' are inaccurate—specifically, overly pessimistic—and that these inaccuracies exacerbate intergroup conflict. Through a review of existing work and a reanalysis of published data, we (i) develop a typology of the disparate phenomena that are labelled 'polarization', (ii) use that typology to distinguish actual from (mis)perceived polarization, and (iii) identify when misperceived polarization gives rise to actual polarization (e.g. extreme issue attitudes and prejudice). We further suggest that a specific psychological domain is ideal for developing corrective interventions: *meta-perception*, one's judgement of how they are perceived by others. We review evidence indicating that correcting meta-perception inaccuracies is effective at reducing intergroup conflict and discuss methods for precisely measuring meta-perception accuracy. We argue that the reputational nature of meta-perception provides a motivational mechanism by which individuals are sensitive to the truth, even when those truths pertain to the 'other side'. We conclude by discussing how these insights can be integrated into existing research seeking to understand polarization and its negative consequences.

This article is part of the theme issue 'The political brain: neurocognitive and computational mechanisms'.

1. Introduction

Concerns about political polarization and its negative effects on democracy and intergroup relations have increased among scholars, policy makers and the public alike over the past few years [1–4]. Yet despite this attention to polarization as a major contributor to modern political ills, a growing body of scientific work on 'false polarization' has simultaneously flourished, suggesting that the extent of polarization is largely a fiction of our minds [5–8].

So are we actually quite polarized or do we just *think* we are? The answer, ultimately, is it depends. Of course, there is a temptation when we think about polarization and its negative consequences to champion interpartisan harmony, but a well-functioning democracy *requires* disagreement and debate. Thus, our aim is to promote belief accuracy regarding the true levels of polarization and the true attitudes of outgroups. However, before we get to that we need to disentangle the myriad phenomena that are labelled 'polarization' and understand their psychological roots.

Here, we argue that integrating the psychology of meta-cognition—beliefs about what other people believe—will help us better understand polarization. First, we construct a typology of psychological and intergroup phenomena that, while distinct, have all been labelled 'polarization'. With this typology, we argue that what is commonly called 'false polarization' is best understood as *inaccurate* meta-cognitive beliefs of the first-order (what they believe) and second-order (what they believe about us). However, the term 'false polarization' is misleading: polarization is a real, measurable phenomenon in the world. As we will review, what is false is people's *beliefs about the extent* of that polarization.

Table 1. Phenomena called 'polarization'.

	level of analysis			
	actual polarization objective attributes of people and intergroup relations		(mis)perceived polarization subjective perceptions of people and intergroup relations (meta-cognition)	
	individual positions	intergroup gaps	first-order beliefs	second-order beliefs (meta-perception)
ideology	issue-position extremity	the actual 'Partisan Gap' on issues, ideological polarization	perceived outgroup extremism, assumed disagreement, perceptions of the 'Partisan Gap'	what I think they think I believe: ideological meta-perceptions, felt misunderstanding
identity	partisan identification extremity	actual bimodal distribution of identification	how I think other people identify with their party, my belief about the size of the party identification gap	how I think they think I identify: meta-stereotypes
outgroup feelings, attributions	prejudice, dehumanization of outgroups, affective polarization	actual tribalism, true intergroup animosity	negative motive attributions, assumed distrust	how I think they feel about us/our actions: group meta-perceptions, meta-dehumanization

To avoid confusion and facilitate conceptual clarity, we argue that the term 'false polarization' should be replaced with 'misperceived polarization'.

We then highlight a growing body of work indicating that inaccurate perceptions of polarization can nonetheless drive actual polarization. To bolster our claim that inaccurate meta-cognition is a significant driver of actual polarization, we conduct a novel reanalysis of published data on intergroup meta-cognitive biases to demonstrate how psychological methods for measuring judgement accuracy can be used to better understand misperceived polarization and its relationship to actual polarization. We conclude by arguing that interventions designed to target inaccurate meta-perceptions may be particularly effective at reducing inaccurate perceptions of polarization and their negative effects owing to the reputationally relevant nature of second-order beliefs (i.e. that they are about 'me' or 'us').

(a) A typology of polarization, actual and (mis)perceived

As research in this area has flourished across multiple social science disciplines, many distinct phenomena across differing levels of analysis have been labelled 'polarization'. For example, polarization, specifically affective polarization, is often operationalized as the extremity of an individual's prejudice towards a political outgroup [3,9,10]. Other times, polarization is defined as individuals' ideological or issue-position extremity [11,12], or as the strength of their partisan ingroup identification [13]. Sometimes, polarization is defined as these phenomena at an intergroup rather than the individual level of analysis. That is, polarization is the *empirical gap* in ideology, outgroup attitudes or ingroup identification between parties [14–16].

Yet, the third class of polarization phenomena refer to individuals' *beliefs about* outgroup individuals' positions or the intergroup gaps [1,7,8]. For example, there are several documented cases in which people harbour overly negative beliefs about what the 'other side' believes [5,17,18]. The term 'false polarization' seems to have arisen from research on 'naive realism' showing that partisans overestimate disagreement with outgroups (see [19,20]).

While many individual papers make these distinctions between different types of polarization (e.g. 'actual versus perceived', 'affective versus ideological'), the absence of an organizing framework not only makes the synthesis of the literature difficult, it makes discussing 'false polarization' all the more confusing. In order for something to be 'false', there must be a corresponding true value to which it is compared. To distinguish between polarization which relates to objective features of people and the world (actual polarization), and polarization that is a subjective perception which may or may not be accurate (perceived polarization), we propose a typology of phenomena called 'polarization' in table 1, which delineates between four levels of analysis and three domains of psychological processes.

In table 1, individual and intergroup phenomena, such as one's own political issue position or outgroup prejudice, are *objective* features of individuals' psychology and social relations, respectively. Meta-cognitive beliefs, by contrast, are *subjective* perceptions of how I see 'them' or the gaps between us (first-order beliefs), and perceptions of how we think 'they' see 'us' (second-order beliefs; what we call meta-perceptions). The critical insight here is that (i) so-called false polarization is solely the domain of meta-cognitive beliefs (the right two columns), and (ii) the distinction between 'false' and actual polarization is more a distinction between levels of analysis (states of the world/people versus

perceptions of those states), not between phenomena that do or do not exist.

We also break these phenomena down by three psychological processes along the rows in table 1. Ideological processes relate to phenomena that are issue-oriented, including extremism, naive realism and perceived intergroup disagreement. Identity processes relate to the strength of identification with social groups (e.g. political parties) for the self, fellow ingroup members and the outgroup. Outgroup feelings and attributions are phenomena that are generally other-focused, domain/issue-general and affective, including prejudice, distrust and assumed negative reciprocity.

(b) How inaccuracy in first versus second-order beliefs lead to actual polarization

Table 1 highlights the manner in which (mis)perceived polarization is conceptually distinct from other forms of polarization, but also how (mis)perceived polarization can refer to two distinct meta-cognitive judgements: inaccurate first-order beliefs about how others identify or how far apart two groups are on an issue (e.g. perceived partisan gap), and inaccurate second-order beliefs about what others think about *oneself* and *one's group* (meta-perceptions). This distinction not only helps us integrate disparate findings across the literature, but it also begins to illuminate the way in which inaccuracies in first versus second-order judgements may arise from different mechanisms.

(i) First-order beliefs

Understanding inaccurate first-order beliefs is paramount, as such inaccuracies have been linked to a host of negative intergroup outcomes. For example, *perceived* out-party polarization (measured as perceived outgroup policy attitude extremity) is more strongly associated with negative outgroup evaluations than is perceiver's level of *actual* polarization (their own policy preferences; [14]). Similarly, partisans vastly overestimate the levels of party-stereotypic membership (e.g. the percentage of Democrats who are lesbian, gay, bisexual, or transgender); these inaccurate perceptions correlate with out-party dislike and participants' feelings of social distance from out-party members [21]. Partisans also underestimate how much they agree with out-party member views, which drives overestimation of the negative affect associated with exposure to opposing views and reduced consumption of opposing views [22].

Several scholars have begun to investigate whether correcting first-order belief inaccuracy can attenuate intergroup animus. A growing body of work conducted after the 2016 US President Election has found that fact-checking is successful at increasing belief accuracy [23–25]; however, many such interventions focus on correcting factual beliefs about the world/public policy (e.g. beliefs about crime statistics) rather than inaccurate meta-cognitive beliefs. Evidence that correcting such *factual* beliefs reduces actual polarization is weak, because fact-checking is less effective when the corrective information is directly counter-attitudinal (i.e. debunking personal ideology) relative to when it is unrelated to ideology [26]. By contrast, addressing first-order misperceptions regarding polarization appears to be promising. For example, an informational intervention correcting inaccurate perceptions of out-party policy extremity actually reduced participants' own attitude extremity [27], providing direct evidence that correcting inaccurate first-order beliefs can reduce actual polarization.

What drives first-order belief (in)accuracy? The evidence is conflicting. For example, in judging the political views of other individuals, those with *more* extreme political attitudes are in fact more accurate judges [28]. By contrast, when inaccuracies are examined among perceptions of the intergroup *gap* in ideological positions, those with *less* extreme partisan positions [17], weaker ingroup identity [14,21] and less political sophistication [1] are more accurate. These disparate findings highlight the need for scholars to carefully disambiguate the inaccurate beliefs they are investigating, as inaccuracies across differing phenomena (e.g. factual statistics versus the attitudes of specific others versus the true intergroup gap in attitudes) may have distinct psychological antecedents.

(ii) Second-order beliefs

Negative and inaccurate second-order beliefs about how 'they' see 'us' (meta-perceptions) also play a central role in driving intergroup conflict and have been noted as a likely contributor to 'toxic' polarization [4]. For example, Democrats and Republicans with the most extreme ideological attitudes were the most likely to overestimate the levels of prejudice and dehumanization their respective out-party held towards them. These inaccurate meta-perceptions were, in turn, uniquely associated with a willingness to violate democratic norms in favour of ingroup loyalty [29]. Inaccurate meta-perceptions also play a role outside the domain of politics. For example, the belief that one's social group is dehumanized by the other group incites reciprocal dehumanization and support for hostile actions against outgroup members [30]. Similarly, racial meta-stereotypes (what are 'their' stereotypes about my group?) are associated with anxiety and decreased self-esteem [31,32].

What drives meta-perception (in)accuracy? While research on first-order belief inaccuracy has tended to focus on individual explanations (e.g. information deficits, attitude extremity), research on inaccurate meta-perceptions has generally found that features of the intergroup contexts (e.g. whether the groups are in competition), rather than individual characteristics, are a strong predictor of (in)accuracy. For example, while Democrats and Republicans exhibit equally inaccurate group meta-perceptions in competitive contexts, reframing the same intergroup interactions as cooperative yields accurate meta-perceptions across both parties [33]. Among Israelis and Palestinians, group meta-perception accuracy was associated with perceived political losses/gains after a conflict incident (a feature of the intergroup context), but not by participants' political knowledge, left-right orientation or empathy for the outgroup [34]. Finally, the belief that one is dehumanized by outgroups (meta-dehumanization) is independent of one's prejudice towards the outgroup [30].

These intergroup context-contingent patterns of meta-perception accuracy parallel research on dyadic meta-perceptions, where the relationships between people are generally better predictors of accuracy than attributes of the individuals. For example, meta-perceptions become significantly less accurate in competitive versus cooperative work contexts [35], paralleling the findings from [33]. Also, in relationships research, the nature of the relationship (i.e. strangers, friends or romantic partners) is a much stronger predictor of meta-perceptive accuracy than how much individuals like each other or perceive the relationship to be of high quality [36,37].

In summary, inaccurate second-order beliefs (meta-perception) represent a unique vector by which misperceived polarization leads to actual polarization via mutual reinforcement, distinct from the mechanisms associated with inaccurate first-order judgements. Therefore, understanding the methods psychologists have long used to measure meta-perception, and disentangle its many psychological components, presents a generative avenue for scholars interested in studying polarization.

(c) Measuring inaccurate meta-perception: a reanalysis

If social scientists are to combat the cycle between misperceived polarization and actual polarization, they must be able to carefully measure the specific meta-cognitive beliefs that are inaccurate and connect those inaccuracies to downstream consequences of interest. To highlight the power of existing psychological methods for measuring meta-cognitive accuracy, and disentangling the multiple components of judgement (in)accuracy, we turn to decades of research on the accuracy of meta-perception.

Meta-perception accuracy has long been of interest to psychology [38–41]. Questions of meta-accuracy have been studied largely at the individual level, such as meta-perception in the domain of personality judgements [42], close relationships [36], the workplace [35] and stereotyping [32]. From early writings on judgement accuracy [39], through the social relations model [43] to more recent methodologies like the social accuracy model [44] and the truth and bias model [45], scholars have acknowledged that 'accuracy' in social perception in truth has multiple components that can vary independently and have distinct antecedents. For example, meta-perceptions can be decomposed into judgements of how one is uniquely perceived by others (distinctive meta-accuracy), how one is perceived stereotypically (normative meta-accuracy), how one is perceived differently by different observers (differential meta-accuracy), and how one is uniquely misperceived by others (meta-insight). Moreover, these types of meta-accuracy can be operationalized as linear relationships (profile agreement/rank-order accuracy) or mean differences (directional bias).

To demonstrate how integrating some of these distinctions into research on misperceived polarization can help us better understand and combat it, we performed a novel analysis of the data published in [33]. We originally found that group meta-perceptions, second-order beliefs about how one's outgroup perceives the collective behaviour of the ingroup, were highly negative and inaccurate among Democrats and Republicans. This negativity bias persisted across multiple competitive, but not cooperative, intergroup contexts. Critically, greater inaccuracy was associated with stronger negative first-order beliefs about outgroup motives. We found that an informational intervention informing Democrats and Republican of their inaccurate group meta-perceptions significantly reduced negative motive attributions towards the outgroup and was more effective on partisans who exhibited greater baseline inaccuracy.

Note that we examined accuracy as a matter of mean differences: are group meta-perceptions higher or lower than the average actual perception of outgroup members? Nonetheless, the structure of the data in our experiment 4 allows for a componential (re)analysis of group meta-perceptive accuracy. Using the social accuracy model [44], we can examine accuracy as a matter of mean over/underestimation within judgement, and across judgements as the linear relationship between perceptions and their respective true values. Critically, this analysis will reveal patterns of meta-perceptive (in)accuracy not detailed in the original paper and provide further insight into our understanding of meta-perception biases and their relationship to actual polarization.

2. Methods

Below we conduct a novel reanalysis of the data from experiment 4 of [33]. We preregistered the original collection of these data on the Open Science Framework (OSF) (https://osf.io/atck5); the data and analysis code for the published results are publicly available (https://osf.io/zhysa/). We have also made our reanalysis of the original data available online (https://osf.io/4z6rc/).

(a) Sample

The sample of 536 participants consisted of self-identified Democrats and Republicans in the United States who participated in the experiment in March 2019. Participants were recruited through Qualtrics survey panels and were quota matched to census distributions along the following variables to ensure the sample was nationally representative: age, gender, ethnicity, education and income (see [33] supplementary materials for exact quotas). The sample also had a quota for a 50/50 split of Democrats and Republicans. No participants who completed the survey are excluded from data analysis.

(b) Procedure

Participants were randomly assigned, between-subjects, to one of three conditions: the actual perception, ingroup perception or group meta-perception condition. Within the condition, participants read five scenarios in randomized order, and for each scenario responded to three items. In the actual perception condition, the scenarios pertained to the participant's outgroup acting competitively towards their ingroup, and participants answered how much *they* disliked, opposed and found politically unacceptable the behaviours in the scenarios. In the ingroup perception condition, participants received the same stimuli as the actual perception condition but were asked the perceptual items at the level of the 'average' *ingroup* member rather than their own perceptions. In the group meta-perception condition, the scenarios pertained to the participant's ingroup acting competitively against their outgroup, and the items ask participants to judge how the average *outgroup* member would perceive the behaviours in the scenarios. Across all conditions, all perception measurements used 0–100 unipolar sliding scales. Exact survey materials can be found on the OSF here: (https://osf.io/pbeaz/).

(c) Analysis

Lees & Cikara [33] analysed accuracy as the mean difference between responses in each condition (i.e. group meta-perception inaccuracy equalled the difference between the group meta-perception and actual perception conditions). Here, we adopted a componential approach where accuracy was understood as both mean level over/underestimation of a given value, and the within-participant linear rank-order relationship between participant's judgements and their respective true values [44,45]. This captures (i) point-estimate accuracy within each scenario/judgement (e.g. accuracy operationalized as whether each the group meta-perception rating is equal to its corresponding actual perception among the outgroup), and (ii) rank-order accuracy across all scenarios/judgements (e.g. accuracy operationalized as whether the relative ranking of the group meta-perceptions across all

judgements match the order of actual perceptions made by outgroup members).

For participants in the group meta-perception condition, the true values are the mean responses from outgroup members, within item and scenario, in the actual perception condition. For participants in the ingroup perception condition, the true values are the mean responses from ingroup members, within item and scenario, in the actual perception condition.

Drawing from the social accuracy model [44], we used a unified linear mixed-effect model framework to analyse accuracy in participant judgements. Participants' judgements across the group meta- and ingroup perception conditions were modelled as the dependent variable, while within-participant true values (actual-perception condition means), condition (0 = ingroup perceptions, 1 = group meta-perceptions), participant party affiliation (0 = Democrat, 1 = Republican), and the three-way interaction between them were modelled as predictors. Random intercepts for participant, scenario, and the interaction between participant and scenario were also modelled. *p*-values and degrees of freedom were calculated using Satterthwaite approximation. Both the true values and participant judgements were centred on the grand mean of the true values [45,46], which orthogonalized the variables and allowed the intercept to reflect the mean difference between judgements and the truth (i.e. conceptually replicating the original analysis from [33]).

(d) Results

Table 2 presents the results from the analysis of the linear relationship between perceptions and the true values of how perceivers' respective in- and outgroup actually perceived the behaviour in the scenarios. In line with the findings from [33], the intercept estimate 9.26 ($p < 0.001$) reproduced the mean differences between the actual and ingroup perceptions conditions originally observed, and the condition fixed-effect estimate 11.38 ($p < 0.001$) reproduced the mean difference between the group meta- and ingroup perceptions conditions observed in the original study.

What is new is that we observed a significant linear relationship between participant judgements and the true values, suggesting that participants were accurate about the beliefs of the groups they were forecasting; however, this relationship was qualified by significant two- and three-way interactions. As such, we calculated marginal slope estimates for this linear relationship by party and condition. Figure 1 visualizes the results. Democrats ($b = 0.71$, 95% confidence interval (CI) = (0.53, 0.90)) and Republican ($b = 0.50$, 95% CI = (0.27, 0.73)) in the ingroup perceptions condition exhibited relative rank-order accuracy in judging the true perceptions of members of their ingroup. However, in the group meta-perception condition, we observed no relative rank-order accuracy among Democrats ($b = -0.03$, 95% CI = (−0.24, 0.19)) or Republican ($b = 0.16$, 95% CI = (−0.02, 0.34)) in judging the true perceptions of members of the outgroup.

Put simply, the difference in accuracy for our judgements for the ingroup and our meta-perception judgements for the outgroup was not just a matter of degree, but a matter of *kind*. When estimating our ingroup's beliefs we are still sensitive to the varying severity of the scenarios and get the rank-order of judgements right; the same is not true when we are considering the beliefs of the outgroup.

(e) Discussion

These results highlight the critical importance, and benefit, of using a componential approach to understanding accuracy in meta-cognitive judgement. Here, we replicated the original finding that group meta-perceptions are, at a mean level, overly pessimistic for forecasts of both ingroup and outgroup members. However, when examining accuracy as the linear relationship between perceptions and true values, we found that participants were accurate in their rank-order estimates of fellow ingroup members' beliefs, but not outgroup members' beliefs, despite both Democrats and Republicans having near-identical actual-perceptions of the scenarios.

Rather than concluding that partisans overestimate polarization when forecasting the perceptions of all other partisans (with merely an attenuated effect size when comparing perceptions of the ingroup versus outgroup), this reanalysis suggests a qualitatively different interpretation: partisans misperceive outgroup polarization, but *accurately* perceive ingroup polarization, with some upward bias in their mean estimates. This more nuanced interpretation is owing to the use of componential analyses of accuracy and highlights the utility of incorporating such methods into research on (mis)perceived polarization.

This pattern also suggests that overly negative group meta-perceptions do not result from a lack of knowledge *per se*, rather that they result from an inability (or unwillingness) to apply the knowledge individuals already possess—in this case, the relative perceived extremity of the scenarios. Such interpretations broadly align with our assertion that misperceived polarization in the domain of second-order judgements is largely a factor of intergroup context rather than individual differences or information deficits. Thus one implication of these findings is that the links between inaccurate judgement and actual polarization reviewed above arise from a particular kind of inaccuracy: one in which one's model of the 'other' is severely over-generalized. While we may have strong stereotypes about both the other party and our

Table 2. Accuracy of ingroup and group meta-perception. (Statistically significant *p*-values less than 0.05 are in bold.)

	Judgement		
predictors	*b*	95% CI	*p*
intercept	9.26	4.54–13.98	**<0.001**
true values	0.71	0.53–0.90	**<0.001**
condition [meta-P]	11.38	5.92–16.83	**<0.001**
party [Republican]	0.99	−4.83–6.81	0.739
true values * condition [meta-P]	−0.74	−0.96–−0.52	**<0.001**
true values * party [Republican]	−0.21	−0.45–0.03	0.082
condition [meta-P] * party [Republican]	−0.41	−8.27–7.45	0.919
(true values * condition [meta-P]) * party [Republican]	0.40	0.08–0.72	**0.015**
random effects			
σ^2	173.06		
$\tau_{00\ \text{Scenario:ID}}$	452.44		
$\tau_{00\ \text{ID}}$	237.52		
$\tau_{00\ \text{Scenario}}$	6.98		
ICC	0.80		
N_{Scenario}	5		
N_{ID}	366		
observations	5479		
marginal R^2/ conditional R^2	0.070/0.815		

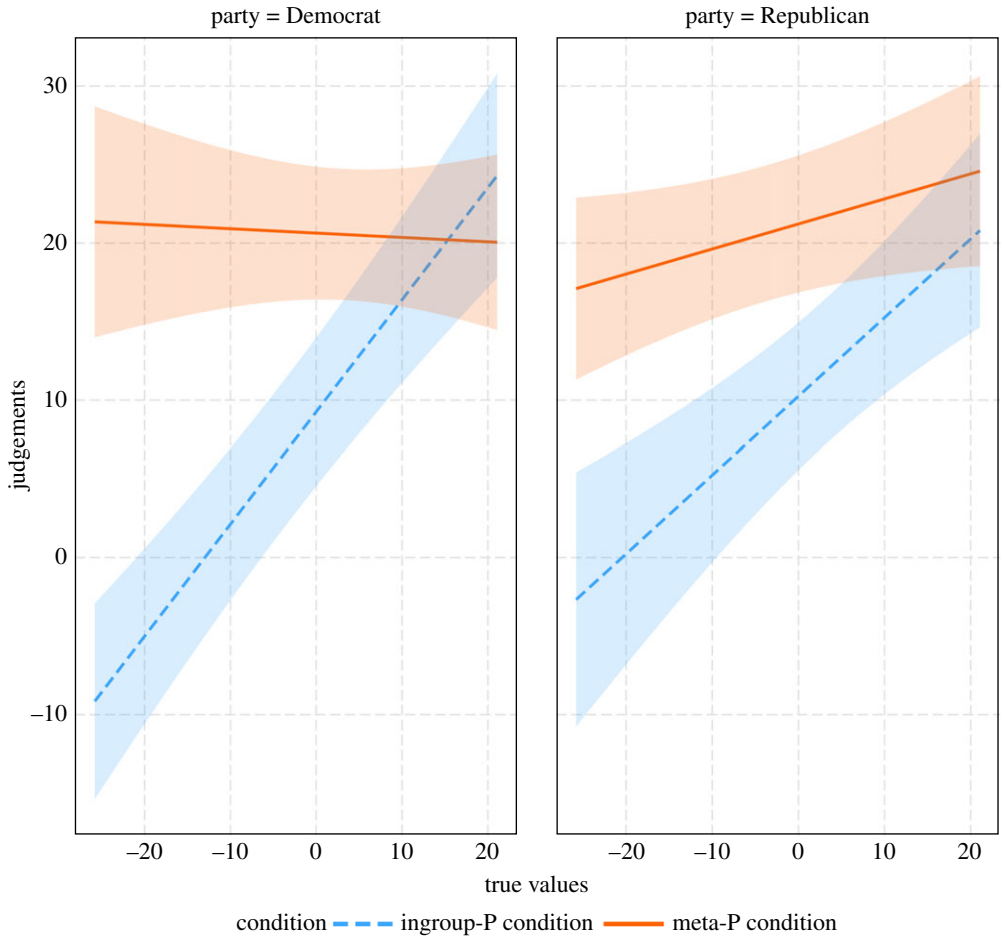

Figure 1. $N = 366$ (N observations = 5479). Plot of the three-way interaction between party-identification (by the panel), condition: ingroup (dashed line) versus meta-perception (solid line), and other participants' 'true' values in predicting participants' judgements. For both parties, results indicate significant and positive linear relationships between ingroup perceptions and their true values, providing evidence for rank-related accuracy; we find no such evidence for rank-related accuracy in group meta-perceptions. Error bars 95% CIs. (Online version in colour.)

own, our beliefs about 'us' are at least still modulated by the details of the local context. Examining how partisan extremity affects the accuracy of ingroup meta-perception is a fruitful avenue of future research, as recent work suggests ideological extremity exacerbates the inaccuracy of outgroup meta-perception among Democrats and Republicans [29].

3. Focus on correcting inaccurate meta-perceptions

Integrating research on meta-cognitive inaccuracy into our understanding of polarization provides two useful insights for scholars developing interventions to correct inaccurate beliefs. The first insight is the need for componential analyses of inaccuracy. Researchers have long noted the univariate measures of judgement accuracy can be misleading [39,44,47,48] as they not only collapse across accuracy's multiple components, such as rank-order and mean-level accuracy, they also fail to account for known biases that affect such judgements, including normativity and stereotyping [49], and projection [50]. If we wish to uncover the mechanisms driving misperceived polarization and use those mechanisms to develop interventions to increase belief accuracy, we need to have a precise picture of the nature of such inaccuracy.

The second insight is that informational interventions for reducing belief inaccuracy may be more effective on inaccurate second-order beliefs than on first-order beliefs. While both first- and second-order beliefs are subject to the motivation to perceive the world accurately, second-order beliefs are unique in that they are fundamentally tied to one's own, or one's groups, reputation. Individuals have a strong motivation to manage the positive impression they make on others [51,52], and in order to do so, they need to accurately understand how they are perceived by others. This reputational motive might explain why meta-perceptions are more accurate in cooperative contexts [33,35] and among those in closer relationships [36]. While work on correcting inaccurate meta-perceptions in intergroup relations is still nascent, we predict that the reputational nature of meta-perception makes individuals uniquely sensitive to corrective information in a way that first-order beliefs and factual beliefs may not.

More broadly, we encourage polarization scholars to consider inaccurate meta-cognitive beliefs more centrality in their models of polarization, a call others have recently made as well [4]. Perhaps understandably, existing work on inaccurate political beliefs tends to focus on inaccurate factual beliefs (e.g. is climate change real, was Barack Obama born in the United States), yet interventions attempting to update such inaccurate beliefs tend to have small effects on judgement accuracy and struggle to overcome entrenched beliefs (see meta-analysis in [26]). While there is still much research to be done, recent work suggests corrective interventions targeting inaccurate meta-cognitive beliefs are quite effective at

reducing negative intergroup outcomes [22,30,33]. Such interventions present a fruitful avenue for attenuating the mutually reinforcing relationship between misperceived and actual polarization.

4. Conclusion

Here, we argue that so-called false polarization is best understood as inaccurate first- and second-order beliefs in intergroup contexts, distinct in kind from actual perception, and better referred to as '(mis)perceived' polarization. We explicate the theoretical and empirical use of this distinction by disambiguating first-order and second-order meta-cognitive beliefs, highlighting existing work on inaccuracies in these domains, describing the methods used for carefully measuring inaccurate meta-cognitive beliefs and conducting a novel reanalysis of existing work on inaccurate group meta-perceptions. We believe that integrating these insights into research on polarization will help social scientists develop more effective interventions for breaking the mutually reinforcing cycle between inaccurate intergroup beliefs and negative political outcomes.

Authors' note. Parts of the 'Measuring inaccurate meta-perception: a re-analysis' section are based on unpublished portions of Jeffrey Lees' dissertation, and parts of the 'Sample' and 'Procedure' sections are based on the text in [33].

Ethics. Experiment 4 of [33] was approved by Harvard University's Institutional Review Board, and all participants gave their informed consent before participating.

Data accessibility. Data and analysis scripts for the findings presented here can be found at https://osf.io/4z6rc/.

Authors' contributions. All authors contributed substantially to the conceptualization and drafting of this article, and approved the final version. J.L. conducted the data analysis presented herein.

Competing interests. We declare we have no competing interests.

Funding. Work on this project by M.C. was supported by a National Science Foundation Award (no. BCS-1551559).

References

1. Armaly MT, Enders AM. In press. The role of affective orientations in promoting perceived polarization. *Political Sci. Res. Methods*. (doi:10.1017/psrm.2020.24)
2. Federico CM. In press. When do psychological differences predict political differences? Engagement and the psychological bases of political polarization. In *Political polarization* (ed. J-W vanProoijen). London, UK: Routledge.
3. Iyengar S, Lelkes Y, Levendusky M, Malhotra N, Westwood SJ. 2019 The origins and consequences of affective polarization in the United States. *Ann. Rev. Political Sci.* **22**, 129–146. (doi:10.1146/annurev-polisci-051117-073034)
4. Moore-Berg SL, Hameiri B, Bruneau E. 2020 The prime psychological suspects of toxic political polarization. *Curr. Opin. Behav. Sci.* **34**, 199–204. (doi:10.1016/j.cobeha.2020.05.001)
5. Blatz CW, Mercier B. 2018 False polarization and false moderation: political opponents overestimate the extremity of each other's ideologies but underestimate each other's certainty. *Soc. Psychol. Pers. Sci.* **9**, 521–529. (doi:10.1177/1948550617712034)
6. Druckman JN, Klar S, Krupnikov Y, Levendusky M, Ryan JB. 2019 The illusion of affective polarization. See https://www.ipr.northwestern.edu/documents/working-papers/2019/wp-19-25.pdf.
7. Levendusky MS, Malhotra N. 2016 (Mis)perceptions of partisan polarization in the American public. *Public Opin. Q.* **80**(S1), 378–391. (doi:10.1093/poq/nfv045)
8. Westfall J, Van Boven L, Chambers JR, Judd CM. 2015 Perceiving political polarization in the United States: party identity strength and attitude extremity exacerbate the perceived partisan divide. *Perspect. Psychol. Sci.* **10**, 145–158. (doi:10.1177/1745691615569849)
9. Simas EN, Clifford S, Kirkland JH. 2020 How empathic concern fuels political polarization. *Am. Political Sci. Rev.* **114**, 258–269. (doi:10.1017/S0003055419000534)
10. Stone DF. 2020 Just a big misunderstanding? Bias and Bayesian affective polarization. *Int. Econ. Rev.* **61**, 189–217. (doi:10.1111/iere.12421)
11. Rollwage M, Dolan RJ, Fleming SM. 2018 Metacognitive failure as a feature of those holding radical beliefs. *Curr. Biol.* **28**, 4014–4021.e8. (doi:10.1016/j.cub.2018.10.053)
12. Stanley ML, Henne P, Yang BW, De Brigard F. 2019 Resistance to position change, motivated reasoning, and polarization. *Political Behav.* **42**, 891–913. (doi:10.1007/s11109-019-09526-z)
13. Mason L. 2018 Ideologues without issues: the polarizing consequences of ideological identities. *Public Opin. Q.* **82**(S1), 866–887. (doi:10.1093/poq/nfy005)
14. Enders AM, Armaly MT. 2018 The differential effects of actual and perceived polarization. *Political Behav.* **41**, 815–839. (doi:10.1007/s11109-018-9476-2)
15. Navajas J, Álvarez Heduan F, Garrido JM, Gonzalez PA, Garbulsky G, Ariely D, Sigman M. 2019 Reaching consensus in polarized moral debates. *Curr. Biol.* **29**, 4124–4129.e6. (doi:10.1016/j.cub.2019.10.018)
16. Porter T, Schumann K. 2018 Intellectual humility and openness to the opposing view. *Self Identity* **17**, 139–162. (doi:10.1080/15298868.2017.1361861)
17. Van Boven L, Judd CM, Sherman DK. 2012 Political polarization projection: social projection of partisan attitude extremity and attitudinal processes. *J. Pers. Soc. Psychol.* **103**, 84–100. (doi:10.1037/a0028145)
18. Yang J *et al.* 2016 Why are 'others' so polarized? Perceived political polarization and media use in 10 countries. *J. Comput.-Mediated Commun.* **21**, 349–367. (doi:10.1111/jcc4.12166)
19. Pronin E, Pucio C, Ross L. 2002 Understanding misunderstanding: social psychological perspectives. In *Heuristics and biases: the psychology of intuitive judgment* (eds T. Gilovich, D. W. Griffin, D. Kahneman), pp. 636–665. Cambridge, UK: Cambridge University Press.
20. Robinson RJ, Keltner D, Ward A, Ross L. 1995 Actual *versus* assumed differences in construal: 'naive realism' in intergroup perception and conflict. *J. Pers. Soc. Psychol.* **68**, 404–417. (doi:10.1037/0022-3514.68.3.404)
21. Ahler DJ, Sood G. 2018 The parties in our heads: misperceptions about party composition and their consequences. *J. Politics* **80**, 964–981. (doi:10.1086/697253)
22. Dorison CA, Minson JA, Rogers T. 2019 Selective exposure partly relies on faulty affective forecasts. *Cognition* **188**, 98–107. (doi:10.1016/j.cognition.2019.02.010)
23. Nyhan B, Porter E, Reifler J, Wood TJ. 2020 Taking fact-checks literally but not seriously? The effects of journalistic fact-checking on factual beliefs and candidate favorability. *Political Behav.* **42**, 939–960. (doi:10.1007/s11109-019-09528-x)
24. Porter E, Wood TJ, Kirby D. 2018 Sex trafficking, Russian infiltration, birth certificates, and pedophilia: a survey experiment correcting fake news. *J. Exp. Political Sci.* **5**, 159–164. (doi:10.1017/XPS.2017.32)
25. Wood T, Porter E. 2019 The elusive backfire effect: mass attitudes' steadfast factual adherence. *Political Behav.* **41**, 135–163. (doi:10.1007/s11109-018-9443-y)
26. Walter N, Cohen J, Holbert RL, Morag Y. 2020 Fact-checking: a meta-analysis of what works and for whom. *Political Commun.* **37**, 350–375. (doi:10.1080/10584609.2019.1668894)
27. Ahler DJ. 2014 Self-fulfilling misperceptions of public polarization. *J. Politics* **76**, 607–620. (doi:10.1017/S0022381614000085)
28. Ivanov I, Muller D, Delmas F, Wänke M. 2018 Interpersonal accuracy in a political context is moderated by the extremity of one's political

attitudes. *J. Exp. Soc. Psychol.* **79**, 95–106. (doi:10.1016/j.jesp.2018.07.001)

29. Moore-Berg SL, Ankori-Karlinsky L-O, Hameiri B, Bruneau E. 2020 Exaggerated meta-perceptions predict intergroup hostility between American political partisans. *Proc. Natl Acad. Sci. USA* **117**, 14 864–14 872. (doi:10.1073/pnas.2001263117)
30. Kteily N, Hodson G, Bruneau E. 2016 They see us as less than human: metadehumanization predicts intergroup conflict via reciprocal dehumanization. *J. Pers. Soc. Psychol.* **110**, 343–370. (doi:10.1037/pspa0000044)
31. Finchilescu G. 2010 Intergroup anxiety in interracial interaction: the role of prejudice and metastereotypes. *J. Soc. Issues* **66**, 334–351. (doi:10.1111/j.1540-4560.2010.01648.x)
32. Vorauer JD, Hunter A, Main K, Roy S. 2000 Meta-stereotype activation: evidence from indirect measures for specific evaluative concerns experienced by members of dominant groups in intergroup interaction. *J. Pers. Soc. Psychol.* **78**, 690–707. (doi:10.1037/0022-3514.78.4.690)
33. Lees J, Cikara M. 2020 Inaccurate group meta-perceptions drive negative out-group attributions in competitive contexts. *Nat. Hum. Behav.* **4**, 279–286. (doi:10.1038/s41562-019-0766-4)
34. Saguy T, Kteily N. 2011 Inside the opponent's head: perceived losses in group position predict accuracy in metaperceptions between groups. *Psychol. Sci.* **22**, 951–958. (doi:10.1177/0956797611412388)
35. Eisenkraft N, Elfenbein HA, Kopelman S. 2017 We know who likes us, but not who competes against us: dyadic meta-accuracy among work colleagues. *Psychol. Sci.* **28**, 233–241. (doi:10.1177/0956797616679440)
36. Carlson EN. 2016 Meta-accuracy and relationship quality: weighing the costs and benefits of knowing what people really think about you. *J. Pers. Soc. Psychol.* **111**, 250–264. (doi:10.1037/pspp0000107)
37. Carlson EN, Furr RM. 2009 Evidence of differential meta-accuracy: people understand the different impressions they make. *Psychol. Sci.* **20**, 1033–1039. (doi:10.1111/j.1467-9280.2009.02409.x)
38. Carlson EN, Vazire S, Furr RM. 2011 Meta-insight: do people really know how others see them? *J. Pers. Soc. Psychol.* **101**, 831–846. (doi:10.1037/a0024297)
39. Cronbach LJ. 1955 Processes affecting scores on 'understanding of others' and 'assumed similarity'. *Psychol. Bull.* **52**, 177–193. (doi:10.1037/h0044919)
40. Kenny DA, DePaulo BM. 1993 Do people know how others view them? An empirical and theoretical account. *Psychol. Bull.* **114**, 145–161. (doi:10.1037/0033-2909.114.1.145)
41. Laing RD, Phillipson H, Lee AR. 1966 *Interpersonal perception: a theory and method of research*. Berlin, Germany: Springer.
42. Vazire S. 2010 Who knows what about a person? The self-other knowledge asymmetry (SOKA) model. *J. Pers. Soc. Psychol.* **98**, 281–300. (doi:10.1037/a0017908)
43. Kenny DA, Albright L. 1987 Accuracy in interpersonal perception: a social relations analysis. *Psychol. Bull.* **102**, 390–402. (doi:10.1037/0033-2909.102.3.390)
44. Biesanz JC. 2010 The social accuracy model of interpersonal perception: assessing individual differences in perceptive and expressive accuracy. *Multivariate Behav. Res.* **45**, 853–885. (doi:10.1080/00273171.2010.519262)
45. West TV, Kenny DA. 2011 The truth and bias model of judgment. *Psychol. Rev.* **118**, 357–378. (doi:10.1037/a0022936)
46. Enders CK, Tofighi D. 2007 Centering predictor variables in cross-sectional multilevel models: a new look at an old issue. *Psychol. Methods* **12**, 121–138. (doi:10.1037/1082-989X.12.2.121)
47. Barranti M, Carlson EN, Côté S. 2017 How to test questions about similarity in personality and social psychology research: description and empirical demonstration of response surface analysis. *Soc. Psychol. Pers. Sci.* **8**, 465–475. (doi:10.1177/1948550617698204)
48. Wood D, Furr RM. 2016 The correlates of similarity estimates are often misleadingly positive: the nature and scope of the problem, and some solutions. *Pers. Soc. Psychol. Rev.* **20**, 79–99. (doi:10.1177/1088868315581119)
49. Furr RM. 2008 A framework for profile similarity: integrating similarity, normativeness, and distinctiveness. *J. Pers.* **76**, 1267–1316. (doi:10.1111/j.1467-6494.2008.00521.x)
50. Ames DR. 2004 Strategies for social inference: a similarity contingency model of projection and stereotyping in attribute prevalence estimates. *J. Pers. Soc. Psychol.* **87**, 573–585. (doi:10.1037/0022-3514.87.5.573)
51. Bolino MC, Kacmar KM, Turnley WH, Gilstrap JB. 2008 A multi-level review of impression management motives and behaviors. *J. Manage.* **34**, 1080–1109. (doi:10.1177/0149206308324325)
52. Leary MR, Kowalski RM. 1990 Impression management: a literature review and two-component model. *Psychol. Bull.* **107**(I), 34–47. (doi:10.1037/0033-2909.107.1.34)