PHILOSOPHICAL TRANSACTIONS B

royalsocietypublishing.org/journal/rstb

Review

Cite this article: Lau T. 2021 Reframing social categorization as latent structure learning for understanding political behaviour. *Phil. Trans. R. Soc. B* **376**: 20200136.
https://doi.org/10.1098/rstb.2020.0136

Accepted: 17 November 2020

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

Subject Areas:
cognition, neuroscience, behaviour

Keywords:
latent structure learning, social categorization, social groups

Author for correspondence:
Tatiana Lau
e-mail: tatiana.lau@rhul.ac.uk

Electronic supplementary material is available online at https://doi.org/10.6084/m9.figshare.c.5303115.

THE ROYAL SOCIETY
PUBLISHING

Reframing social categorization as latent structure learning for understanding political behaviour

Tatiana Lau

Department of Psychology, Royal Holloway, University of London, Egham TW20 0EX, UK

TL, 0000-0002-0681-7295

Affiliating with political parties, voting and building coalitions all contribute to the functioning of our political systems. One core component of this is social categorization—being able to recognize others as fellow in-group members or members of the out-group. Without this capacity, we would be unable to coordinate with in-group members or avoid out-group members. Past research in social psychology and cognitive neuroscience examining social categorization has suggested that one way to identify in-group members may be to directly compute the similarity between oneself and the target (dyadic similarity). This model, however, does not account for the fact that the group membership brought to bear is context-dependent. This review argues that a more comprehensive understanding of how we build representations of social categories (and the subsequent impact on our behaviours) must first expand our conceptualization of social categorization beyond simple dyadic similarity. Furthermore, a generalizable account of social categorization must also provide domain-general, quantitative predictions for us to test hypotheses about social categorization. Here, we introduce an alternative model—one in which we infer latent groups of people through latent structure learning. We examine experimental evidence for this account and discuss potential implications for understanding the political mind.

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

1. Introduction

Our ability to affiliate with groups, choose allies and build coalitions can be linked to the functioning and maintenance of political institutions, elections and parties. Yet, these same abilities can also contribute to parochialism and political polarization, which have been growing in recent years. For example, a text analysis of historical US congressional speeches found that polarization (operationalized as the ability of an observer to correctly identify the political party of the speaker based on the text) has reached an all-time high [1]. Political tensions have likewise reached new highs; for example, people report levels of bias along political party lines that are similar to levels reported along racial lines [2].

One crucial precedent to cooperating in groups and organizing into political movements is the ability to sort ourselves and others into in- and out-group members, or social categorization. This is a fundamental capacity for group living [3], and a better understanding of the mechanisms underlying this capacity has potential implications for understanding political preferences, voting behaviour and how to combat current political ills. This review offers a critical discussion of the current social categorization literature with an emphasis on recent advances that conceptualize social categories as latent groups and social categorization as latent structure learning. We conclude by discussing future directions and potential applications of latent structure learning in the context of understanding political behaviour.

2. Social categorization

In our interconnected social world where we encounter new people all the time, how do we accurately categorize others and choose our allies? One possibility suggested by social psychology research is to rely on dyadic similarity—we could compare our own social identities against those of the target person to infer similarity to the target, whether through direct or implied means (e.g. assigned team membership and audio-visual cues). This could be done through explicit signals; for a Trump supporter, seeing someone wear a 'Make America Great Again' cap would be an indicator of sharing a social group membership with the target person (i.e. being Trump supporters). We can also rely on other visual cues such as skin tone and gender to infer group membership (e.g. [4]). Other studies have demonstrated how children rely on familiar accents to choose between social targets [5]. Many other studies of intergroup relations have directly sorted participants into explicit, minimal teams to demonstrate in-group favouritism [6]. Even the most contrived, assigned, explicit team memberships can trigger intergroup bias and conflict, as Sherif [7] demonstrated in his seminal Robbers Cave study, where boys from similar backgrounds who were unfamiliar with one another were separated into groups at a summer camp; after they bonded within their groups, the introduction of competitive camp activities resulted in general intergroup aggression (e.g. name-calling, flag burning, stealing, etc.).

Of course, similarity need not be inferred along a single dimension or absolute. Our social selves contain a multitude of social group memberships, and the combination of one's group memberships can also be important. For example, research in intersectionality highlights how racial identity and its impact on one's welfare cannot be considered independently of gender and class [8,9]. Particular mixes of different social group memberships (e.g. being African-American and female) can lead to specific adverse outcomes [10]. Children express strong pro-White biases when presented with Black and White faces, but these biases manifest at different levels for males and females [11]. Furthermore, people can also perceive gradations of similarity. For example, religious children perceive a religious character (albeit one of a different religion) as more similar to themselves than a non-religious character, and this similarity correlates with reported liking for the character [12]. Dyadic similarity based on gradation can also be found in political science theories such as simple spatial voting models, wherein voters support candidates with whom their policy preferences directly align the most [13]. Across all of these, calculations of relationships are dyadic in nature; how one perceives another person is affected only by the degree of similarity between oneself and the target person. Strong considerations of dyadic similarity along salient dimensions may drive social preference and inference of group memberships.

Yet, even though dyadic similarity may remain the same across time and space, our understanding of in- and out-group membership is mutable; we have the ability to reassemble and reorganize into superordinate groups and different coalitions when necessary [14,15]. Indeed, the group membership(s) we bring to bear in any given situation is (are) context-dependent [16]. Additionally, whereas past studies of group membership have typically provided some form of accessible cues to group membership (e.g. [17]), we do not always have direct access to other people's group memberships. Even if there are visual cues to group membership, these may not be relevant to the dimension along which we sort others into coalitional in- and out-group members. Given our ability to flexibly reshape 'us' and 'them' as per the context of the situation and without needing explicit visual cues, a simple account of dyadic similarity does not suffice to account for our diverse social behaviours.

Previous theories further highlight potential examples of how context affects social categorization. Some accounts, such as optimal distinctiveness theory, state that specific social identities are made salient through a balance of resolving the inner conflict of needing to both belong to a group and maintain individuality [18], and these social identities might be the dimensions along which we divide the world. Thus, dyadic similarity between us and a target person may not change across different contexts, but being among people whose group memberships highlights one particular dimension on which we and the target differ may lead us to group the target with the out-group, while a different composition of others may highlight dimensions that we and the target share and make us regard the target as an in-group member. Additionally, Campbell [19] identified sharing a common fate, or the degree of shared outcomes among group members, in addition to similarity as core components of entitativity. Shared fate may serve as a contextual effect. Sherif [7] also demonstrated the importance of functional relations between groups (i.e. whether or not the groups were in competition or in cooperation) as an input to social categorization. This is captured in self-categorization theory, which posits that contexts and environments play a role in social categorization; someone whom one may infer as a fellow group member in one context may not be considered an in-group member in another context [16]. It may not be a person's similarities on preferences, traits and characteristics that are crucial to inferring that person's social group in relation to one's own, but rather the interaction between the immediate context and the dyadic similarity along a dimension made salient by the context that allow social group inference. Yet, these accounts still sidestep any description of the exact computations allowing for the flexibility with which we divide the world into 'us' and 'them'.

These exact computations are important because a generalizable account of social categorization needs to provide quantifiable predictions that can then be tested. While theories such as those highlighted in the previous paragraph *qualitatively* describe the context-dependence of social categorization, they proffer no *quantitative* predictions about the effects of particular contexts on the resulting categorizations. In other words, a computational account of social categorization would allow us to input quantitative 'settings' (e.g. the degree to which people under consideration are similar to us, the number of people under consideration, etc.) and generate quantitative predictions. This is required for us to develop concrete hypotheses about when and why people perceive a target as an in-group member in one context but not the next. While dyadic similarity, in a particular form, can provide quantitative predictions (e.g. summing the number of similarities between oneself and a target), these calculations are dyadic in nature and predict the same outcome in every context. By contrast, a generalizable computational model would allow us to produce predictions about the degree of similarity between targets and perceivers that is required for

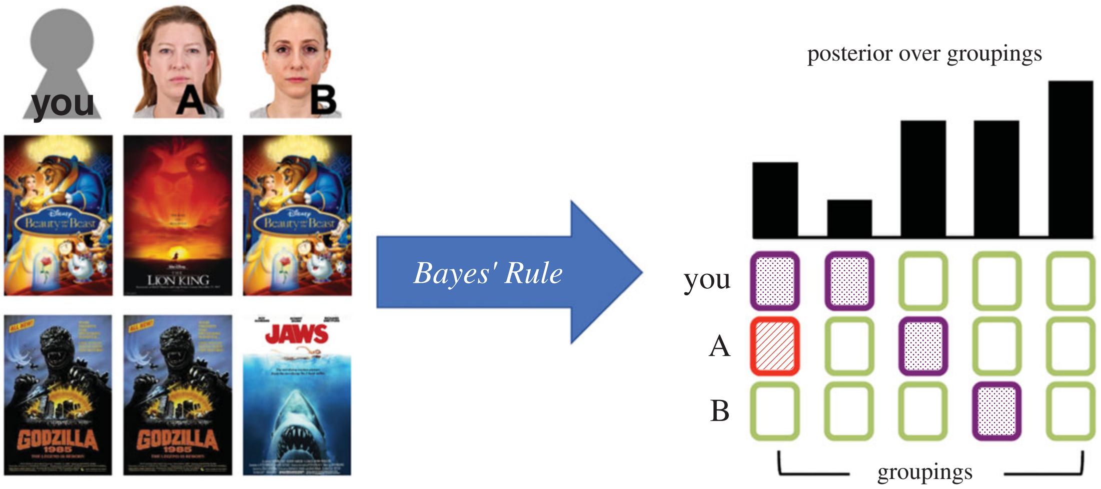

Figure 1. A schematic of latent structure learning. Columns of coloured boxes on the bottom right represent all possible group configurations given three people, ranging from all three being in separate groups (indicated by the three differently coloured/patterned boxes) to all three people being in one group (indicated by all three people being assigned to one colour/pattern). We can use observed behaviour (e.g. the movie preferences of each person) to update a prior in order to generate a posterior distribution over all possible groupings and infer the most probable latent grouping of people. In this case, the most likely grouping is one where all three people are all in one group together. (Online version in colour.)

perceivers to view targets as in-group members. Finally, a computational account would allow us to model group affiliation and social influence across larger social networks. Without a computational formalization of our theories, we fail to understand mechanisms and the exact contexts under which the in-group or out-group divisions can occur.

3. Latent structure learning

Rather than thinking purely in terms of categorical social groups, we could abstract away and instead start from the literature on statistical processes through which groups or clusters of data points can be determined [20]; under these processes, observations sharing more features or characteristics with one another are assigned to the same cluster. While some cluster analysis methods require us to predefine the number of clusters (e.g. *k*-means clustering [21]), others infer clusters by allowing for probability distributions over the different possible distributions of the data (e.g. Dirichlet processes [22]). If it is the case that we cluster social others and ourselves in a similar manner to how we might statistically determine clusters in observed data points, then accounts of social group inference incorporating these computations may help us both infer social groups and simultaneously determine the salient dimension(s) along which we draw social divisions.

Applying this to social categorization, we may be able to infer latent groups, or clusters of people [23,24]. Under this account, we can use our observations of others' choice patterns to infer what the most probable group configuration of people is across multiple individuals (figure 1). This can be determined using Bayes' Rule, in which our posterior (the group configuration given the choice patterns) is the product of a prior (the probability of the group configurations) and a likelihood (the choice patterns given the group configuration). We can set our prior using a Chinese restaurant process [25]. This clustering method is analogous to customers walking into a Chinese restaurant with an infinite seating capacity and tables; in this case, our tables are analogous to groups and customers are analogous to choices. The probability of a new customer coming in and sitting down at a particular table is a function of the number of people already seated at that table and a dispersion parameter. This dispersion parameter governs the probability that the new customer will be seated at an empty table (i.e. as it approaches infinity, each person will be assigned to their own table for one). While there can be an infinite number of tables in theory, a 'rich get richer' dynamic will favour more parsimonious groupings [26].

We can then compare these possible group configurations with the observed choice patterns; our likelihood will dictate that the more similar the observed choice behaviour is to a hypothesized grouping, the higher the probability assigned to that grouping. In other words, if two people, Annie and Betsy, always choose the same movies, while you and a third person, Carol, always choose the same set of movies, albeit one that is different from the set that Annie and Betsy choose, then the most likely group configuration is one wherein you are grouped with Carol and Annie is paired with Betsy. If all four of you were to then choose four different movies, this observation would weaken the probability assigned to this particular group configuration. By inferring a probability distribution over all possible latent groupings given the number of people under consideration, this model allows us to infer an infinite number of groups over an infinite number of people but eliminates the need for predetermining (i) the number of social groups and (ii) the dimensions along which one carves the relevant social groups.

It is important to note that this account does not necessitate the existence of one or more out-group(s). For example, if everyone is identical in his/her choices (e.g. everyone, including you, supports the same taxation rates), the model would assign the highest probability to a grouping where everyone is assigned to the same group as you. At the other extreme, if each person is distinct enough (e.g. each person prefers a very different set of taxation policies), then the generative model would assign the highest probability to the grouping where each person is assigned to his/her own group. Thus, given the right context, a perceiver may identify everyone as a fellow member of the in-group.

Recent studies have designed a task to test this account [27]. Participants were asked in a task to learn about the preferences of new people. Participants stated their own

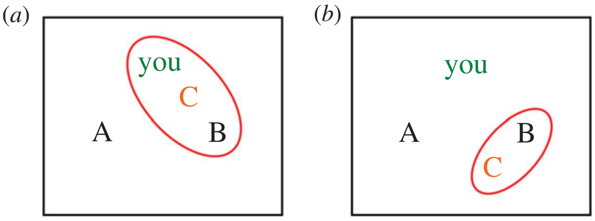

Figure 2. An abstract representation of similarity space. Represented are two scenarios in which you are equally similar to Annie and Betsy, and Carol is more similar to Betsy than to Annie. (*a*) Carol is very similar to you, so you infer a latent structure consisting of yourself, Carol and Betsy. (*b*) Carol is not very similar to you, so you infer a latent structure consisting of only Carol and Betsy (you are not in the group). (Online version in colour.)

preference ('yes' or 'no') on a political issue (e.g. 'Should genetically modified foods be labelled?') and then guessed and learned through feedback the preferences of three other people on that same issue. After doing this for a series of different issues, participants reached a test trial where they had to choose between two of the three people's 'mystery' choices. Participants were told that these two people had informed experimenters of their preferences on a new political issue, but the participant was privy to neither the issue nor the people's chosen stances. Given that participants had just learned about the preferences of these two people on a series of other political issues, with whom would the participant rather ally? In effect, participants were being asked to make an 'ally-choice' and ally with one of the two people given their knowledge of these people's preferences on the previous political issues.

This particular task structure effectively allows one to pit the predictions of the two models—dyadic similarity and latent structure learning—against each other. At each ally-choice trial, participants ('you' in figure 2) were forced to choose between two people (A and B in figure 2) who were equally similar to them (i.e. each had agreed with the participant on an equal number of political issues). According to the dyadic similarity account, participants should show ambivalence between the two people because they are equally similar to the participant. More importantly, this account posits that participants' choices should not be influenced by the presence of a third person, because the participants' views of A and B are affected only by their direct agreement with A and B, respectively. By contrast, the latent structure learning account would predict that participants' choices between A and B would be affected by the presence of a third person, because the location of this person within the agreement space affects which latent groups are inferred as being most probable. In this particular experimental instantiation of the task, this third person, C, agrees more with B than with A (figure 2). When C agrees a lot with the participant (e.g. agreeing on a majority of the political issues; figure 2*a*), the participant infers a latent group consisting of him/herself, C and B (C's higher rate of agreement with B allows the inclusion of B), and this makes the participant more likely to choose B over A. However, when C agrees very little with the participant (e.g. *disagreeing* on a majority of the political issues; figure 2*b*), the participant should be more likely choose A over B, because the participant infers being left out of the latent group that consists of B and C.

In a series of experiments using a variety of political issues, participants' behaviours reflected the inference of and reliance on these latent groups in order to choose political allies; the presence of C affected how participants chose between A and B [27,28]. Moreover, these latent groups affected trait inferences made about A, B and C. When participants inferred being in the same latent group as B (figure 2*a*), they rated B as more competent, moral and likeable compared with when they inferred being in a separate group from B (figure 2*b*). Furthermore, even when participants could rely solely on dyadic similarity, the latent groups affected their choices. In a final experiment, participants were randomly assigned to one of two coloured teams at the beginning of the experiment. At each ally-choice trial, A and B were shown in coloured boxes, which reflected A's and B's own team memberships. Unbeknownst to the participant, A was always labelled as a member of the same team as the participant, while B was always labelled as a member of the opposite team. If participants stopped relying on latent groups when provided with a clear, easy alternative of explicit dyadic similarity in the form of team memberships, then they should choose A on every ally-choice trial. Instead, participants were still influenced by the degree of agreement between themselves and C in the ally-choice trials. That is not to say that participants never relied on these explicit teams. When participants were in the same latent group as B (the opposite coloured team member; figure 2*a*), they chose B over A about 50% of the time (at chance). However, when explicit team labels and the inferred latent group were in agreement (i.e. when the participant, who was already on the same team as A, inferred not being in the same latent group as B; figure 2*b*), participants were even more likely than in other experiments to side with A. In other words, the explicit groups reinforced the inferred latent groups. Within the latent structure learning framework, we can model the existence of these explicit groups through the dispersion parameter in the prior (e.g. participants may believe *a priori* that there exist only two groups). It is worthwhile noting that these findings are not specific to the political domain; participants exhibit these behavioural effects with cinematic preferences [23], and in large social networks these inferred latent groups are more influential than one's self-reported friendship network for predicting future behaviour [29]. Thus, these experiments and social network analyses demonstrate that latent structure learning may be a more comprehensive behavioural account of social categorization.

4. Neural correlates of social categorization

Recent work extends these accounts, dyadic similarity and latent structure learning, to the brain. How might the brain track social group memberships in the environment? One candidate for dyadic similarity is the medial prefrontal cortex/pregenual anterior cingulate cortex (mPFC/pgACC). Thinking about one's own, and similar others', traits, mental states, perspectives and characteristics elicits activation in this area [30,31]. Additionally, consideration of close others (e.g. family, a group of close friends, etc.) elicits activation in these areas [32]. Tracking dyadic similarity requires understanding others' traits and states relative to one's own to calculate the degree of similarity between oneself and the target. Taken together, this area could compute this difference between others and oneself to track the similarity (or difference) between oneself and another person at any time, making it a good

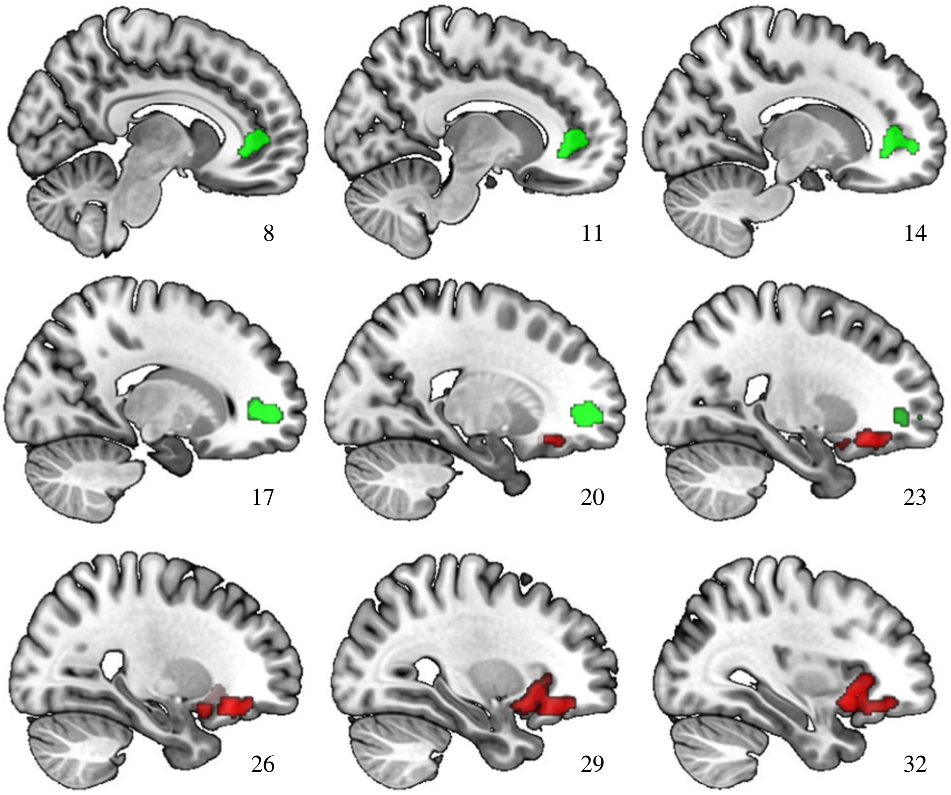

Figure 3. fMRI results from [28]. Whole-brain contrast (family-wise error (FWE)-corrected $p < 0.05$) of parametric modulators: dyadic similarity model (green; pgACC) and latent groups model (red; rAI).

candidate for tracking changes in dyadic similarity. Additionally, understanding the similarities between oneself and the individual under consideration could be a good proxy for whether or not one is in the same group as that person.

On the other hand, tracking latent groups may not require such self-referential processes. A study that asked participants to categorize foods and restaurants to determine which clusters caused illness found that the clustering of causal structures was updated in areas not typically associated with self-referential processes, such as the right anterior insula (rAI; [33]). It is important, however, to note that social categorization is unlike other forms of categorization in that it requires one to additionally classify oneself as a part of the group [16]. For example, one may sort plants into groups of fruits or vegetables without needing to sort oneself into a fruit or vegetable group, but to sort other Democrats and Republicans as in- or out-group members, one must first group oneself as either Democrat or Republican. Differential activity in the rAI has also been linked to categorizing ambiguous race faces and political ideology [34]. In a parcellation of the insula, the anterior insula was found to be connected with the anterior cingulate cortex, amygdala and dorsolateral prefrontal cortex [35]; these areas have also been linked to categorization along racial lines [17,36]. The rAI is well-connected to communicate between areas that would be important for tracking social latent groups.

In an fMRI adaptation of the latent structure learning task described in the previous section, activity in the rAI was correlated with the probability of each individual's membership in the same latent group as the participant [28] (figure 3). Additionally, activity in the pgACC was correlated with tracking and updating dyadic similarity with each individual. Perhaps more importantly, however, the variability in the signal from the rAI, but not the pgACC, was found to improve model predictions of variability in participants' actual choice behaviour on the ally-choice trial. Furthermore, variability in the rAI was much better explained by the latent structure model than by dyadic similarity. This was not true for the pgACC and the dyadic similarity model. Thus, while latent structure learning relies, to some degree, on similarity between people, there is a difference between how the brain tracks dyadic similarity and latent groups. Second, this demonstrates that while the brain simultaneously tracks both dyadic similarity and latent groups, only the variability in the signal from the area tracking the latent groups actually improves predictions of behaviour.

5. The political mind

This latent groups account provides specific, testable, quantitative hypotheses about how different contexts will result in different inferred groupings, which affect whom we perceive as fellow group members and how we choose allies. This model is generalizable across situations and scales with the number of people under consideration. Indeed, if we reframe social categorization as latent structure learning, we can move beyond thinking about political allegiance as a static, immutable affiliation and rather as a changeable, evolving function of all possible current political party allegiances. Taken together, what does this imply for our understanding of political behaviour and organization?

First, the latent groups account has implications for the way in which voting behaviour should be modelled. Common political science accounts of voter behaviour, such as a simple spatial voting model [37], assume that voters will choose the political candidate with whom their preferences directly align the best and these choices should not be influenced by the stances of other voters or the stances of other candidates. Yet, the latent groups account suggests that voting behaviour is much more complex; how voters perceive all candidates will affect how they view any individual candidate because it affects which groups are inferred and which divisions are made relevant. Indeed, it is worth reiterating that only signal variability from the brain region tracking latent groups improved predictions of variability in actual choice behaviour on the ally-choice trial [28]. This demonstrates the importance of latent structures in these ally-choice decisions. The account has implications for how we can interpret recent political events. While the 2016 US Presidential election was ultimately driven by economic factors, sexism and racism [38], the alignment of Hillary Clinton with racial minorities (B and C in figure 2*b*), a group whose views on immigration reform, criminal justice reform, etc. differ from Whites, may have served to help additionally push undecided White voters towards Trump (A in figure 2*b*). In other words, rather than maximizing the number of policy stances on which a candidate and target voters agree, political strategists also need to consider the context under which important undecided voters are viewing the electoral race (e.g. how a candidate is viewed with respect to other voting blocs as well as other candidates matters). More broadly speaking, however, how factions align to form these latent groups should not be ignored, especially when voters are strongly undecided between candidates, and the model may have use for predicting shifts in voting trends.

The account also has implications for how we think about the adoption of political preferences. Investigations of the transmission of political preferences and mobilization (e.g. [39,40]) typically conceptualize social influence as stemming from direct neighbours and friends in a given person's social network. By contrast, the latent groups account focuses on how inferred, hidden groups can potentially exert a more powerful social influence compared with the influences of friends and family [29]. If hidden communities in a social network, rather than a voter's direct social connections, exert more influence on the voter's political preferences and activeness, then strategies to increase transmission of political preferences and mobilization of voters may be better off harnessing the power of these latent groups rather than relying on influence via direct ties.

This is not to say that dyadic similarity accounts and spatial voting models are inaccurate; the neuroimaging results demonstrated that the brain tracks and updates both dyadic similarity and latent groups simultaneously [28]. This is in line with the final experiment in [27], where participants used both explicit labels (i.e. dyadic similarity) and latent groups. Specifically, when the latent group inference and dyadic similarity were aligned to favour one particular person, participants were even more likely to ally with that person. Thus, there is clear potential to leverage both dyadic similarity and latent groups simultaneously in political contexts to derive greater identification with the party and drive voting behaviour.

Furthermore, the latent groups account strengthens our understanding of current levels of perceived polarization. The model assigns higher probabilities to group configurations where similarities within groups are greater. Indeed, recent findings in political science demonstrate that political party identities are increasingly 'sorted' and homogeneous, such that racial, religious and other social identities have become aligned with political party membership [41]. Such a high degree of intragroup similarity would result in high probabilities assigned to the existence of distinct, separate groups, and this may contribute to the meta-perceptions that exacerbate misperceived polarization [42]. Moreover, the model suggests interventions that could allow meaningful interactions between opposing groups. Moderately conservative voters can help serve as the bridge between Democrats and other conservative voters (i.e. serving as C in figure 2*a*), and over time, genuine diversification of the voter bases of political parties will lead to the real possibility of depolarization.

It is important to note that decreasing the homogeneity of political parties need not decrease their coalitional power. Because the latent groups account dictates a context-dependent search for the most probable group, the resulting inferred group can exhibit intragroup cohesion without necessarily exhibiting strong ties. Similarly, sociology research has demonstrated that loosely tied groups can exhibit strong cohesiveness in social networks [43], and political coalitions are no different. Thus, through its properties, the model accounts for effects observed in the political science and sociology literatures and also offers potential paths to alleviating political problems.

Finally, the examples mentioned thus far relate primarily to the domain of Anglo-American politics, where a winner-takes-all system favours a distribution of power across only two political parties. This *status quo* of binary choices in elections may enforce an idiosyncratic notion of the existence of two groups. It is worth restating, however, that the model does not necessitate the existence of only two groups. Thus, this model still has relevance for multi-party democracies such as Germany. It may even be the case that in these coalition-driven democracies, latent groups are more relevant for political strategists to consider.

6. Conclusion

Social categorization underlies many political behaviours, including choosing allies, affiliating with political parties, etc. Through understanding how we sort ourselves into 'us' and 'them', we can gain traction on understanding our social and political world. Here, we introduced evidence for reframing social categorization as a form of latent structure learning and discussed how this latent groups model provides insights into the political mind. We hope that through greater adaptation of this account as a form of social categorization, many fields, including psychology, can move forward in their understanding of the flexible nature of social influence, social groups and social networks.

Ethics. Citation Diversity Statement: Recent work in several fields of science has identified a bias in citation practices such that papers from women and other minority scholars are under-cited relative to the number of such papers in the field. Here I sought to proactively consider choosing references that reflect the diversity of the field in thought, form of contribution, gender, race, ethnicity and other factors. (See electronic supplementary material, Citation Diversity Statement.)

Competing interests. I declare I have no competing interests.

Funding. I received no funding for this study.

Acknowledgements. I am grateful to Mina Cikara, Samuel Gershman, Zachary Ingbretsen and Thomas Pouncy for indispensable collaboration on this work.

References

1. Gentzkow M, Shapiro JM, Taddy M. 2019 Measuring group differences in high-dimensional choices: method and application to congressional speech. *Econometrica* **87**, 1307–1340. (doi:10.3982/ecta16566)
2. Iyengar S, Westwood SJ. 2014 Fear and loathing across party lines: new evidence on group polarization. *Am. J. Polit. Sci.* **59**, 690–707. (doi:10.1111/ajps.12152)
3. Bruner JS. 1957 On perceptual readiness. *Psychol. Rev.* **64**, 123–152. (doi:10.1037/h0043805)
4. Kubota JT, Li J, Bar-David E, Banaji MR, Phelps EA. 2013 The price of racial bias. *Psychol. Sci.* **24**, 2498–2504. (doi:10.1177/0956797613496435)
5. Kinzler KD, Shutts K, DeJesus J, Spelke ES. 2009 Accent trumps race in guiding children's social preferences. *Social Cogn.* **27**, 623–634. (doi:10.1521/soco.2009.27.4.623)
6. Tajfel H, Billig MG, Bundy RP, Flament C. 1971 Social categorization and intergroup behaviour. *Eur. J. Social Psychol.* **1**, 149–178. (doi:10.1002/ejsp.2420010202)
7. Sherif M. 1966 *In common predicament: social psychology of intergroup conflict and cooperation*. Boston, MA: Houghton Mifflin.
8. Crenshaw K. 1991 Mapping the margins: intersectionality, identity politics, and violence against women of color. *Stanford Law Rev.* **43**, 1241. (doi:10.2307/1229039)
9. Rogers LO, Scott MA, Way N. 2014 Racial and gender identity among Black adolescent males: an intersectionality perspective. *Child Dev.* **86**, 407–424. (doi:10.1111/cdev.12303)
10. Sidanius J, Hudson STJ, Davis G, Bergh R. 2018 The theory of gendered prejudice: a social dominance and intersectionalist perspective. In *The Oxford handbook of behavioral political science* (eds A Mintz, L Terris). Oxford, UK: Oxford University Press. (doi:10.1093/oxfordhb/9780190634131.013.11)
11. Perszyk DR, Lei RF, Bodenhausen GV, Richeson JA, Waxman SR. 2019 Bias at the intersection of race and gender: evidence from preschool-aged children. *Dev. Sci.* **22**, e12788. (doi:10.1111/desc.12788)
12. Heiphetz L, Young LL. 2019 Children's and adults' affectionate generosity toward members of different religious groups. *Am. Behav. Scientist* **63**, 1910–1937. (doi:10.1177/0002764219850870)
13. Downs A. 1957 *An economic theory of democracy*. New York, NY: Harper.
14. Gaertner SL, Dovidio JF, Samuel G. 2000 *Reducing intergroup bias: the common in-group identity model*. Philadelphia, PA: Psychology Press.
15. Rand DG, Pfeiffer T, Dreber A, Sheketoff RW, Wernerfelt NC, Benkler Y. 2009 Dynamic remodeling of in-group bias during the 2008 presidential election. *Proc. Natl Acad. Sci. USA* **106**, 6187–6191. (doi:10.1073/pnas.0811552106)
16. Turner JC, Oakes PJ, Haslam SA, McGarty C. 1994 Self and collective: cognition and social context. *Pers. Social Psychol. Bull.* **20**, 454–463. (doi:10.1177/0146167294205002)
17. Kubota JT, Banaji MR, Phelps EA. 2012 The neuroscience of race. *Nat. Neurosci.* **15**, 940–948. (doi:10.1038/nn.3136)
18. Brewer MB. 1991 The social self: on being the same and different at the same time. *Pers. Social Psychol. Bull.* **17**, 475–482. (doi:10.1177/0146167291175001)
19. Campbell DT. 2007 Common fate, similarity, and other indices of the status of aggregates of persons as social entities. *Syst. Res.* **3**, 14–25. (doi:10.1002/bs.3830030103)
20. Zubin J. 1938 A technique for measuring like-mindedness. *J. Abnorm. Social Psychol.* **33**, 508–516. (doi:10.1037/h0055441)
21. MacQueen J. 1967 Some methods for classification and analysis of multivariate observations. In *Proc. 5th Berkeley Symp. Mathematical Statistics and Probability*, vol. 1 (eds LM Le Cam, J Neyman), pp. 281–297. Berkeley, CA: University of California Press.
22. Ferguson TS. 1973 A Bayesian analysis of some nonparametric problems. *Ann. Stat.* **1**, 209–230.
23. Gershman SJ, Pouncy HT, Gweon H. 2017 Learning the structure of social influence. *Cogn. Sci.* **41**, 545–575. (doi:10.1111/cogs.12480)
24. Gershman SJ, Cikara M. 2020 Social-structure learning. *Curr. Dir. Psychol. Sci.* **29**, 460–466. (doi:10.1177/0963721420924481)
25. Aldous DJ. 1985 Exchangeability and related topics. In *École d'Été de Probabilités de Saint-Flour XIII — 1983. Lecture notes in mathematics*, vol. 1117 (ed. PL Hennequin), pp. 1–198. Berlin Heidelberg: Springer. (doi:10.1007/bfb0099421)
26. Gershman SJ, Blei DM. 2012 A tutorial on Bayesian nonparametric models. *J. Math. Psychol.* **56**, 1–12. (doi:10.1016/j.jmp.2011.08.004)
27. Lau T, Pouncy HT, Gershman SJ, Cikara M. 2018 Discovering social groups via latent structure learning. *J. Exp. Psychol. General* **147**, 1881–1891. (doi:10.1037/xge0000470)
28. Lau T, Gershman SJ, Cikara M. 2020 Social structure learning in human anterior insula. *eLife* **9**, e53162. (doi:10.7554/elife.53162)
29. Sowrirajan T, Pentland A, Lau T. Submitted. Distributed inference of latent groups for temporal behavior prediction.
30. Heleven E, Van Overwalle F. 2015 The person within: memory codes for persons and traits using fMRI repetition suppression. *Social Cogn. Affect. Neurosci.* **11**, 159–171. (doi:10.1093/scan/nsv100)
31. Jenkins AC, Macrae CN, Mitchell JP. 2008 Repetition suppression of ventromedial prefrontal activity during judgments of self and others. *Proc. Natl Acad. Sci. USA* **105**, 4507–4512. (doi:10.1073/pnas.0708785105)
32. Krienen FM, Tu P-C, Buckner RL. 2010 Clan mentality: evidence that the medial prefrontal cortex responds to close others. *J. Neurosci.* **30**, 13 906–13 915. (doi:10.1523/jneurosci.2180-10.2010)
33. Tomov MS, Dorfman HM, Gershman SJ. 2018 Neural computations underlying causal structure learning. *J. Neurosci.* **38**, 7143–7157. (doi:10.1523/jneurosci.3336-17.2018)
34. Krosch AR, Jost JT, Van Bavel JJ. 2021 The neural basis of ideological differences in race categorization. *Phil. Trans. R. Soc. B* **376**, 20200139. (doi:10.1098/rstb.2020.0139)
35. Chang LJ, Yarkoni T, Khaw MW, Sanfey AG. 2012 Decoding the role of the insula in human cognition: functional parcellation and large-scale reverse inference. *Cereb. Cortex* **23**, 739–749. (doi:10.1093/cercor/bhs065)
36. Ito TA, Bartholow BD. 2009 The neural correlates of race. *Trends Cogn. Sci.* **13**, 524–531. (doi:10.1016/j.tics.2009.10.002)
37. Adams J, Merrill III S, Zur R. 2020 The spatial voting model. In *The SAGE handbook of research methods in political science and international relations* (eds L Curini, R Franzese), pp. 205–223. Thousand Oaks, CA: Sage Publications.
38. Schaffner BF, Macwilliams M, Nteta T. 2018 Understanding white polarization in the 2016 vote for president: the sobering role of racism and sexism. *Polit. Sci. Q.* **133**, 9–34. (doi:10.1002/polq.12737)
39. Sinclair B. 2011 *The social citizen: peer networks and political behavior*. Chicago, IL: University of Chicago Press.
40. Bond RM, Fariss CJ, Jones JJ, Kramer ADI, Marlow C, Settle JE, Fowler JH. 2012 A 61-million-person experiment in social influence and political mobilization. *Nature* **489**, 295–298. (doi:10.1038/nature11421)
41. Mason L. 2018 *Uncivil agreement: how politics became our identity*. Chicago, IL: University of Chicago Press.
42. Lees J, Cikara M. 2021 Understanding and combating misperceived polarization. *Phil. Trans. R. Soc. B* **376**, 20200143. (doi:10.1098/rstb.2020.0143)
43. Granovetter MS. 1973 The strength of weak ties. *Am. J. Sociol.* **78**, 1360–1380. (doi:10.1086/225469)