CellPress
REVIEWS

but a smaller assortment of hard-to-learn traits, in line with empirical research indicating that languages associated with larger populations show larger vocabularies but simpler grammars. Researchers have also considered how models can bridge the gap between individual-level behavior and population structure to explore cultural differences, showing how more complex societies might exhibit more apparent diversity in personality traits [7] and how denser societies should have weaker norms of fairness [8].

These models are in line with a cultural evolutionary approach that represents, in my view, the most coherent attempt to provide a unified theory of human behavior [9,10]. A science of human behavior must not only integrate individual cognition with social structure but also consider how societal structures emerge, change, and shape information flow. An interdisciplinary approach, informed by theory and empirical data and scaffolded by formal models at multiple levels of analysis, is our best bet.

[1]Department of Cognitive and Information Sciences, University of California, Merced, CA, USA

*Correspondence:
paul.smaldino@gmail.com (P.E. Smaldino).
https://doi.org/10.1016/j.tics.2019.07.005

References

1. Muthukrishna, M. and Schaller, M. (2019) Are collectivistic cultures more prone to rapid transformation? Computational models of cross-cultural differences, social network structure, dynamic social influence, and cultural change. *Pers. Soc. Psychol. Rev.* Published online June 28, 2019. https://doi.org/10.1177/1088868319855783
2. Barnett, G.A. (2011) Mathematical models of the diffusion process. In *The Diffusion of Innovations: A Communication Science Perspective* (Vishwanath, A. and Barnett, G.A., eds), pp. 103–122, Peter Lang
3. Hindersin, L. and Traulsen, A. (2014) Counterintuitive properties of the fixation time in network-structured populations. *J. R. Soc. Interface* 11, 20140606
4. Watts, D.J. (2011) *Everything Is Obvious (Once You Know the Answer)*, Crown Business
5. De, S. *et al.* (2018) Tipping points for norm change in human cultures. In *Social, Cultural, and Behavioral Modeling. SBP-BRiMS 2018*In *Lecture Notes in Computer Science* (Vol. 10 899) (Thomson, R. *et al.*, eds), Springer
6. Reali, F. *et al.* (2018) Simpler grammar, larger vocabulary: how population size affects language. *Proc. Biol. Sci.* 285, 20172586
7. Smaldino, P.E. *et al.* Niche diversity can explain cross-cultural differences in personality structure. *Nat. Hum. Behav.* (in press)
8. Schank, J.C. *et al.* (2015) Evolution of fairness in the dictator game by multilevel selection. *J. Theor. Biol.* 382, 64–73
9. Mesoudi, A. (2011) *Cultural Evolution: How Darwinian Theory Can Explain Human Culture and Synthesize the Social Sciences*, University of Chicago Press
10. Henrich, J. (2017) *The Secret of Our Success: How Culture Is Driving Human Evolution, Domesticating Our Species, and Making Us Smarter*, Princeton University Press

Forum

What Underlies Political Polarization? A Manifesto for Computational Political Psychology

Max Rollwage,[1,2,*]
Leor Zmigrod,[3,4] Lee de-Wit,[3]
Raymond J. Dolan,[1,2] and
Stephen M. Fleming[1,2]

Polarization is one of the biggest societal challenges of our time, yet its drivers are poorly understood. Here we propose a novel approach – computational political psychology – which uses behavioral tasks in combination with formal computational models to identify candidate cognitive processes underpinning susceptibility to polarized beliefs about political and societal issues.

Polarization of opinions and beliefs is a growing feature in countries such as the USA and UK. This divide is often a barrier to constructive discourse between those who adhere to opposing outlooks and is increasingly spilling over into personal distrust and misunderstanding of the 'other' side [1]. As this development threatens open societies, it is crucial to understand the mechanisms underpinning the polarization of beliefs about political and societal issues, exemplified by controversies surrounding the UK's EU Referendum and attitudes towards climate change.

One profitable approach in political psychology is to identify 'cognitive styles' – content-free styles of thinking – that are linked to specific political ideologies (see [2] for a comprehensive review). An initial wave of findings has enabled researchers to sketch out a conceptual landscape that maps cognition onto politics; for instance, revealing a link between conservative worldviews and intolerance of uncertainty and need for order and structure [2]. However, in a majority of studies the definition of cognitive styles remains qualitative in nature, operationalized by subjective self-reports from questionnaires, with considerable variability in definition [2]. This renders it difficult to critically appraise and unify existing findings to identify cognitive processes supporting the development of polarized beliefs.

Here we advocate a new approach that involves the use of behavioral tasks in conjunction with formal computational models to uncover an algorithmic basis for cognitive styles. Computational models formalize algorithmic solutions to solve behavioral tasks where different models specify different ways in which information is processed. We suggest that well-validated behavioral tasks (informed by findings in cognitive neuroscience) can reveal differences in computational model parameters and enable the discovery of candidate neural processes from which distinct cognitive styles may emerge. As an example of this approach, a model of Bayesian belief updating describes the normative combination of previous knowledge with new

CellPress
REVIEWS

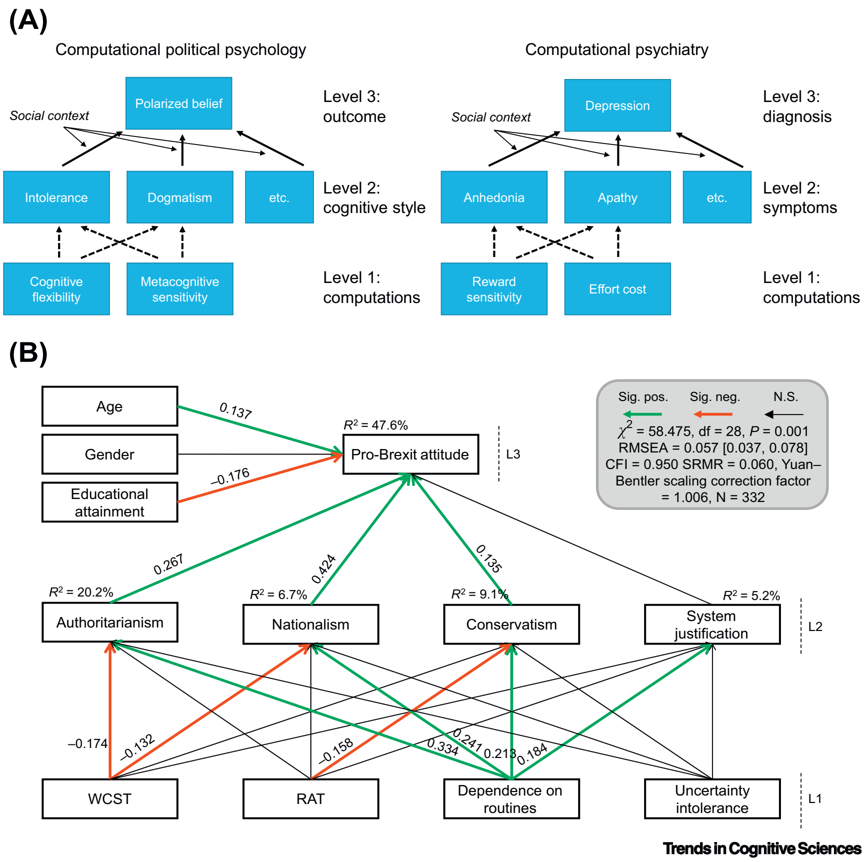

Figure 1. Computational Political Psychology. (A) Illustration of how core computations affecting belief formation could give rise to variation in cognitive styles, which in turn shape polarized views. The left panel shows the proposed schematic for computational political psychology while the right panel shows the analogous approach in computational psychiatry. In both areas, the lowest level is formed by alterations in core computations, which can be measured by using behavioral tasks in combination with computational modeling. Changes in core computations give rise to the next level in the hierarchy (cognitive styles or symptoms). The broken arrows indicate that those links between computations and the second level in the hierarchy represent hypotheses that await empirical testing. Here, the main goal is to understand computational mechanisms giving rise to level 2. Finally, cognitive styles (or symptoms) might shape specific polarized beliefs (or mental health diagnoses) in interaction with social and environmental factors. (B) Example study linking different levels of analysis [9]. Structural equation model predicting support for Brexit: cognitive inflexibility on the Wisconsin Card Sorting Test (WCST) and the Remote Associates Test (RAT), as well as heightened dependence on daily routines, predicts elevated authoritarianism, conservatism, and nationalism, which in turn predict support for Brexit in the UK's 2016 EU Referendum. All parameters shown are fully standardized and significant parameter estimates are shown in green and red bolded lines. Significance level was $P < 0.05$. L1, level 1 (psychological flexibility variables); L2, level 2 (ideological orientation variables); L3, level 3 (attitude outcome variable); N.S., not significant; Sig. neg., significant negative pathway; Sig. pos., significant positive pathway. Reproduced, with permission, from [9].

information, in which the relative weighting of prior knowledge with new information might differentiate between people with dogmatic and nondogmatic world views.

While earlier research has focused on identifying cognitive styles that differ between people on the left and right sides of the political spectrum, recent efforts have focused on extreme or radical beliefs, which may be particularly relevant for understanding the drivers of polarization [3,4]. Extremism is often defined as the distance of a belief from mainstream opinions [3] and radicalism in terms of how beliefs are held and acted on [5]. While precise definitions vary between researchers, key features of radicalism include a tendency towards extreme/violent actions, strong adherence to ingroup norms, dogmatic beliefs, and intolerance toward opposing views [3,5]. Addressing the cognitive underpinnings of this cluster of behaviors represents a promising approach to understanding the drivers of polarization.

Here, computations required to build internal models of the external environment are of most interest. Evidence accumulation plays a key role in inferring the state of the world to guide our actions, while learning (based on prediction errors) is crucial for updating models in light of these inferences. Alterations in these processes can, in principle, lead to inflexible and intolerant beliefs – key features of a radical mindset. Importantly, these mechanisms are generic to the process of belief formation and independent of the specific belief under consideration.

We propose a general framework for linking these different levels of analysis (Figure 1A, left panel). Core computations supporting belief formation exist at the lowest level. To the extent that alterations in such computational processes help to index stable individual differences, they in turn give rise to variation in cognitive styles such as dogmatism at higher levels. In turn, these cognitive styles, in concert with environmental and social factors, shape the formation and content of (polarized) worldviews.

In recent work we have focused on identifying computational correlates of a subset of features that characterize the radical mindset. It has been previously reported that people with radical and extreme beliefs show overconfidence about political and nonpolitical issues [6,7]. However, one-shot measures of the discrepancy between performance and confidence are unable to disentangle the contributions of confidence bias (a tendency to publicly espouse higher confidence) from changes in metacognitive sensitivity (insight into the correctness of one's beliefs). We have recently employed behavioral tasks (unrelated to politics) in conjunction with computational models to show that confidence alterations in people with dogmatic and intolerant political beliefs are due to reduced insight into the correctness of individual decisions [8]. This study provided initial evidence that domain-general computational differences contribute to cognitive styles, which may in turn predispose people to develop polarized views.

We have also found that reduced cognitive flexibility across multiple cognitive tasks – including alterations in computations supporting set-shifting and reversal learning – was associated with heightened authoritarianism, conservatism, and nationalism, which were in turn predictive of real-world voting behavior and attitudes towards Brexit (Figure 1B; [9]). By relying on nonpolitical tasks to objectively measure cognitive flexibility, this work further illustrates that understanding individual differences in information processing aids an understanding of why people take different positions on highly polarized topics.

The approach we advocate has notable parallels with endeavors known as computational psychiatry (Figure 1A, right panel). After decades of reliance on descriptive diagnostic categories, the field of computational psychiatry now aspires to identify transdiagnostic, and biologically plausible, markers of mental health by combining behavioral assays and computational models [10]. For example, we might hypothesize specific computational changes that give rise to symptoms like anhedonia or apathy, such as where these reflect reduced reward sensitivity and/or inflated effort cost. These specific hypotheses are tested by probing healthy and depressed participants with behavioral tasks, fitting computational models to data to extract latent parameters indexing hypothesized computations, and asking whether these model parameters explain individual differences in associated symptoms. This approach has identified reduced reward sensitivity and increased effort costs as separate subclusters of computational alterations in patients, which may indicate distinct pathophysiological subtypes of depression [11].

Another important parallel with computational psychiatry is the promise of a principled basis for tailoring interventions. Many patients receiving a particular diagnosis fail to respond to treatments, leading to a suspicion that current diagnostic categories do not capture crucial differences in underlying mechanisms. Similarly, by developing a computational approach to radicalism, we can in principle identify appropriate cognitive interventional targets, equipping people with generalizable cognitive skills to process information more accurately and without bias. As a first step in this direction, we have shown that it is possible to enhance domain-general metacognitive sensitivity through cognitive training [12], opening up the possibility that similar training could enable people to better reflect on their beliefs and ameliorate resistance to changes of mind in the face of counterevidence.

In summary, we advocate the use of formal models of computational processes that underlie cognitive styles, which in turn are tightly linked to political and societal attitudes. The promise of this approach is the possibility of moving the field beyond conceptual labels, which are often open to interpretation and debate. While single computational alterations might explain only limited variance in cognitive styles, identifying computational building blocks promises a mechanistic understanding of cognitive styles [9] and may facilitate principled interventions to counteract belief polarization (Box 1). We see this approach – computational political psychology – as building on an extensive body of knowledge about cognitive styles to accelerate a deeper understanding of polarization and political attitudes.

[1]Wellcome Centre for Human Neuroimaging, University College London, London WC1N 3BG, UK
[2]Max Planck University College London Centre for Computational Psychiatry and Ageing Research, London WC1B 5EH, UK
[3]Department of Psychology, University of Cambridge, Cambridge CB2 3EB, UK
[4]Behavioural and Clinical Neuroscience Institute, University of Cambridge, Cambridge CB2 3EB, UK

*Correspondence:
max.rollwage.16@ucl.ac.uk(M. Rollwage).
https://doi.org/10.1016/j.tics.2019.07.006

References

1. Iyengar, S. *et al.* (2019) The origins and consequences of affective polarization in the United States. *Annu. Rev. Polit. Sci.* 22, 129–146
2. Jost, J.T. (2017) Ideological asymmetries and the essence of political psychology. *Polit. Psychol.* 38, 167–208
3. Kruglanski, A.W. *et al.* (2019) Cognitive mechanisms in violent extremism. *Cognition* 188, 116–123
4. van Prooijen, J.W. and Krouwel, A.P. (2019) Psychological features of extreme political ideologies. *Curr. Dir. Psychol. Sci.* 28, 159–163
5. Wintrober, R. (2006) *Rational Extremism: The Political Economy of Radicalism*, Cambridge University Press
6. Toner, K. *et al.* (2013) Feeling superior is a bipartisan issue: extremity (not direction) of political views predicts perceived belief superiority. *Psychol. Sci.* 24, 2454–2462
7. Ortoleva, P. and Snowberg, E. (2015) Overconfidence in political behavior. *Am. Econ. Rev.* 105, 504–535
8. Rollwage, M. *et al.* (2018) Metacognitive failure as a feature of those holding radical beliefs. *Curr. Biol.* 28, 4014–4021
9. Zmigrod, L. *et al.* (2018) Cognitive underpinnings of nationalistic ideology in the context of Brexit. *Proc. Natl. Acad. Sci. U. S. A.* 115, E4532–E4540
10. Montague, P.R. *et al.* (2012) Computational psychiatry. *Trends Cogn. Sci.* 16, 72–80
11. Cooper, J.A. *et al.* (2018) Anhedonia in depression: biological mechanisms and computational models. *Curr. Opin. Behav. Sci.* 22, 128–135
12. Carpenter, J. *et al.* (2019) Domain-general enhancements of metacognitive ability through adaptive training. *J. Exp. Psychol. Gen.* 148, 51–64