What underlies political polarization? A manifesto for computational political psychology

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

[1]Wellcome Centre for Human Neuroimaging, University College London, London WC1N 3BG, United Kingdom

[2]Max Planck University College London Centre for Computational Psychiatry and Ageing Research, London WC1B 5EH, United Kingdom

[3]Department of Psychology, University of Cambridge, Cambridge CB2 3EB, United Kingdom

[4]Behavioural and Clinical Neuroscience Institute, University of Cambridge, Cambridge CB2 3EB, United Kingdom

*Correspondence to: Max Rollwage (max.rollwage.16@ucl.ac.uk)

Keywords

Political psychology, computational modelling, cognitive styles, behavioural tasks, radicalism, polarization

Abstract

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 behavioural tasks in combination with formal computational models in order to identify candidate cognitive processes that underpin susceptibility to polarized beliefs about political and societal issues.

Main text

Polarization of opinions and beliefs is a growing feature in countries such as the US and UK. This divide is often a barrier to constructive discourse between those who adhere to opposing outlooks, and is increasingly spilling over into a personal distrust and a 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 about United Kingdom’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 showing a link of conservative worldviews with intolerance of uncertainty and increased need for order and structure [2]. However, in a majority of studies the definition of cognitive styles remains qualitative in nature, operationalised by subjective self-reports from questionnaires, with considerable variability in their definition between studies [2]. This renders it difficult to critically appraise and unify existing findings in order to identify cognitive processes supporting the development of specific beliefs.

Here we advocate a new approach that involves the use of behavioural tasks in conjunction with formal computational models to uncover an algorithmic basis for cognitive styles. Computational models formalise algorithmic solutions to solve a behavioural task where different models specify different ways in which information is processed. We suggest that well-validated behavioural tasks (informed by findings in cognitive neuroscience) can reveal differences in computational model parameters and enable 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 information, in which the relative weighting of prior knowledge with new information might differentiate between people with dogmatic and non-dogmatic world views.

While earlier research has focussed on identifying cognitive styles that differ between people on the left and right sides of the political spectrum, recent efforts have focussed on extreme or radical beliefs, which may be particularly relevant for understanding the drivers of polarisation [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 upon [5]. While precise definitions vary between researchers, key features related to 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 behaviours 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 true state of the world in order to guide our actions, while learning (based on prediction errors) is crucial for updating models in light of these inferences. Failure of these processes can, in principle, lead to inflexible and intolerant beliefs – key features of the 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 of analysis. To the extent that alterations in such computational processes help index stable individual differences, they in turn give rise to variation in cognitive styles such as dogmatism or intolerance 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 non-political 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 behavioural 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 in turn may predispose people to develop polarized views.

Additionally, it was recently found that reduced cognitive flexibility across multiple cognitive tasks – including difficulties with computations supporting set-shifting and reversal learning – was associated with heightened authoritarianism, conservatism, and nationalism, which were in turn predictive of real-world voting behaviour and attitudes towards Brexit ([9], see Figure 1B). By relying on non-political tasks to objectively measure cognitive flexibility, this work further illustrates that understanding individual differences in information processing will ultimately help to understand why people take different positions on highly polarized topics.

The approach we advocate has notable parallels with endeavours known as computational psychiatry (see 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 the combined use of behavioural assays and computational models [10]. For example, we might hypothesise specific computational changes that give rise to symptoms like anhedonia or apathy, such where these reflect reduced reward sensitivity and/or inflated effort cost. These specific hypotheses are tested by probing healthy and depressed participants with behavioural tasks, fitting computational models to data to extract latent parameters indexing hypothesised computations, and asking whether these model parameters explain individual differences in associated symptoms. For instance, this approach has identified reduced reward sensitivity and increased effort costs as separate sub-clusters 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 (see Box 1). 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 a 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 a focus on conceptual labels, which are often open to interpretation and debate. While single computational alterations might only explain 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 in order to accelerate a deeper understanding of polarization and political attitudes.

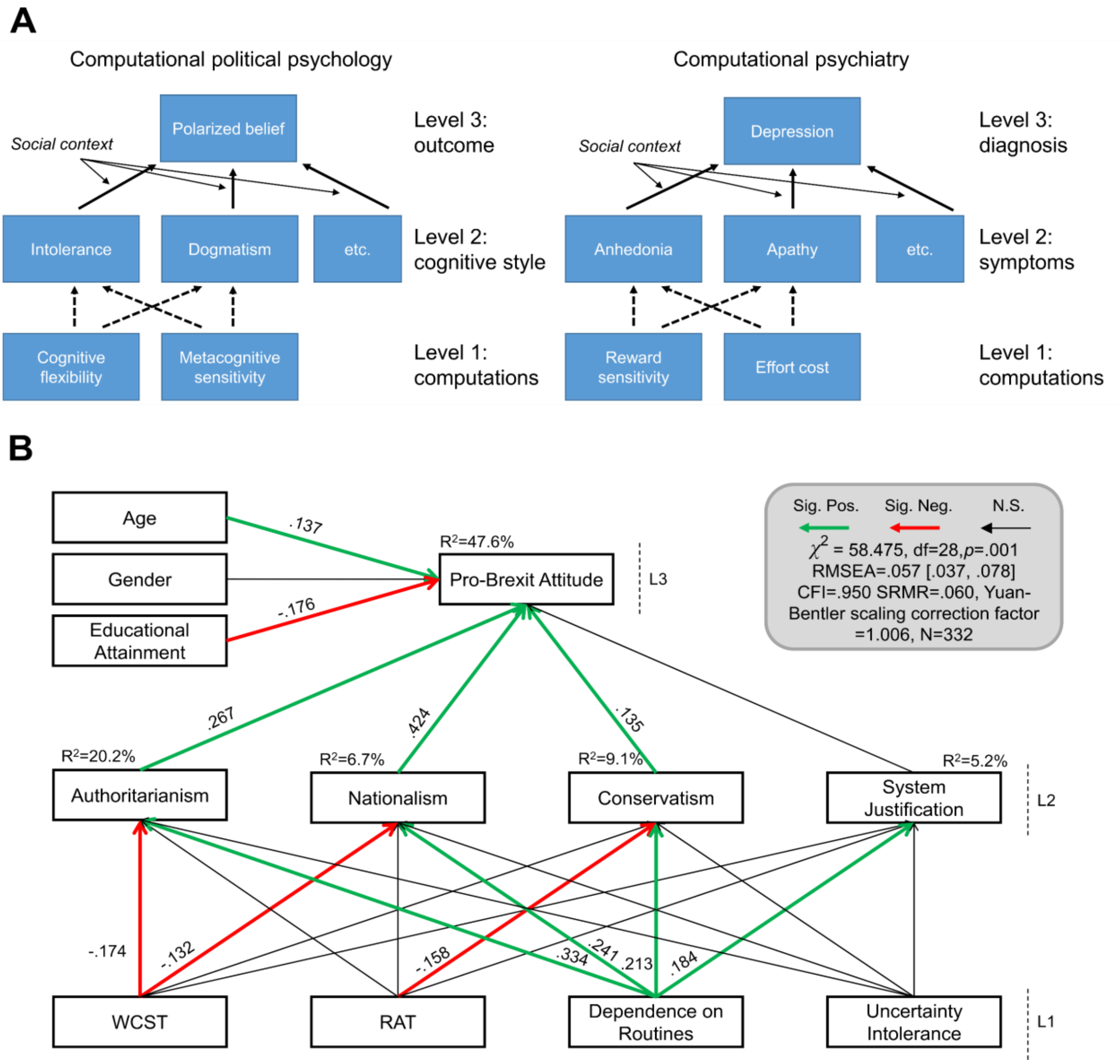

***Figure 1*. The computational political psychology approach. A** Illustration of how core computations affecting belief formation could give rise to variation in cognitive styles which in turn might shape polarized views. The left panel shows the proposed schematic for computational political psychology 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 behavioural tasks in combination with computational modelling. Changes in core computations give rise to the next level in the hierarchy (cognitive styles or symptoms). The dotted 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 the different levels of analysis (Zmigrod et al. 2018). Structural equation model predicting support for Brexit: cognitive inflexibility on the Wisconsin Card Sorting Test (WCST) and Remote Associates Test (RAT), as well as a heightened dependence on daily routines, predict elevated authoritarianism, conservatism, and nationalism, which in turn predict support

for Brexit in the UK's 2016 EU Referendum (Zmigrod et al. 2018). 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 Zmigrod, Rentfrow, & Robbins (2018).

Box 1

The merit of mechanistic understanding

The promise of computational political psychology is in identifying computational building blocks which lead to a mechanistic understanding of cognitive styles. Notably, however, such building blocks may themselves only explain limited variance in political attitudes – as we ourselves have found in recent studies [8]. This is to be expected under the kinds of multilevel models outlined in Figure 1, in which proximal computational mechanisms are related to particular behaviours via changes in personality or symptomology.

For instance, if a doctor attempts to predict whether a person will have a heart attack within 5 years, the best predictor might be the degree to which the coronary arteries contain plaque deposits, with large effect sizes. However, this knowledge does not tell us much about the mechanisms that create plaque deposits and will be unlikely to result in new treatments. Moreover, the contributors to plaque deposits are likely to be multifactorial (e.g. high levels of cholesterol, high blood pressure, etc.) and each of these factors may have only limited influence on the plaque deposit (such that effect sizes for links between individual mechanisms and deposits might be relatively small). Crucially, however, identifying small, reliable effect sizes associated with underlying mechanisms may bring us closer to the possibility of reducing heart diseases through targeted interventions such as a low-cholesterol diet and increased exercise.

Similarly, while cognitive styles may be strong predictors of political behaviour, identifying computational alterations that underpin cognitive styles holds the promise of mechanistic understanding and thus the potential for targeted intervention.

References

1. Iyengar, S., Lelkes, Y., Levendusky, M., Malhotra, N., & Westwood, S. J. (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., Fernandez, J. R., Factor, A. R., & Szumowska, E. (2019) Cognitive mechanisms in violent extremism. Cognition 188, 116-123

4. van Prooijen, J. W., & Krouwel, A. P. (2019) Psychological features of extreme political ideologies. Curr. Dir. in Psychol. Sci. 28, 159-163

5. Wintrober, R. (2006) *Rational extremism: the political economy of radicalism*, Cambridge University Press

6. Toner, K., Leary, M. R., Asher, M. W., & Jongman-Sereno, K. P. (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., & Snowberg, E. (2015) Overconfidence in political behavior. Am. Econ. Rev. 105, 504-35

8. Rollwage, M., Dolan, R. J., & Fleming, S. M. (2018) Metacognitive failure as a feature of those holding radical beliefs. Curr. Biol. 28, 4014-4021

9. Zmigrod, L., Rentfrow, P. J., & Robbins, T. W. (2018) Cognitive underpinnings of nationalistic ideology in the context of Brexit. Proc. Natl. Acad. Sci. 115, E4532-E4540

10. Montague, P. R., Dolan, R. J., Friston, K. J., & Dayan, P. (2012) Computational psychiatry. Trends Cogn. Sci. 16, 72-80

11. Cooper, J. A., Arulpragasam, A. R., & Treadway, M. T. (2018) Anhedonia in depression: biological mechanisms and computational models. Curr. Opinion Behav. Sci. 22, 128-135

12. Carpenter, J., Sherman, M. T., Kievit, R. A., Seth, A. K., Lau, H., & Fleming, S. M. (2019) Domain-general enhancements of metacognitive ability through adaptive training. J. Exp. Psychol. Gen. 148, 51-64