Decoding social knowledge in the human brain

Daniel Alcalá-López[1], David Soto[1,2]

[1]Basque Center on Cognition, Brain and Language (BCBL), Donostia – San Sebastian, Spain
[2]Ikerbasque, Basque Foundation for Science, Bilbao, Spain

d.alcala@bcbl.eu
@danalclop

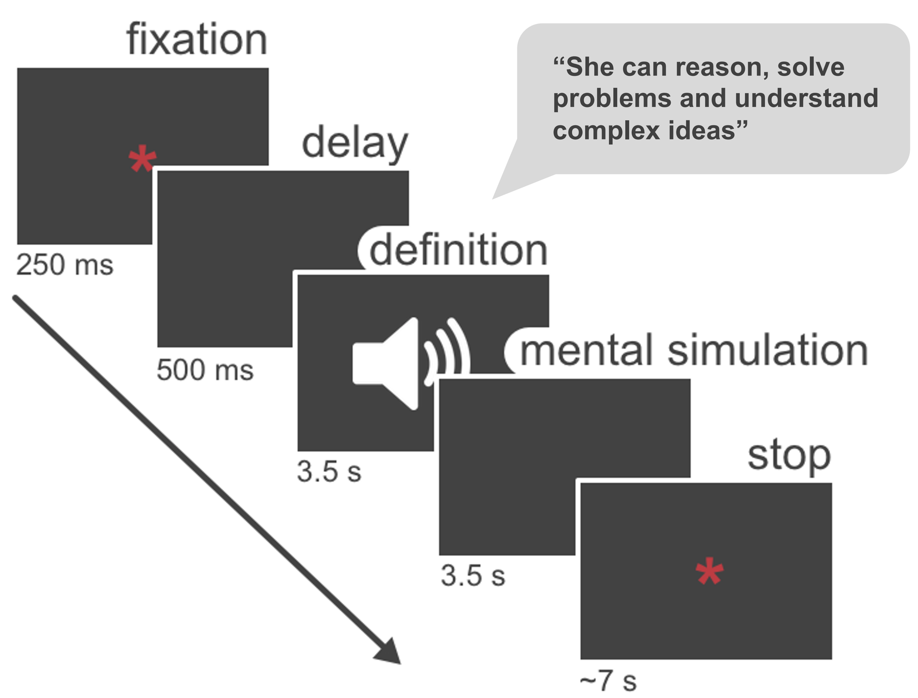
BASQUE CENTER ON COGNITION, BRAIN AND LANGUAGE

Motivation

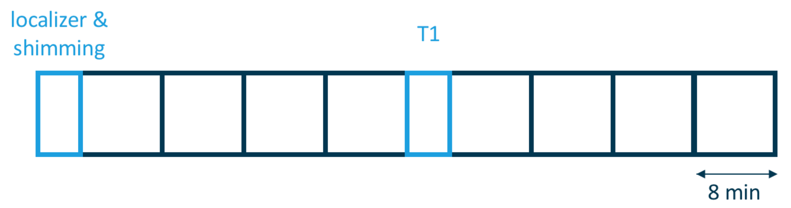
- How does the brain represent conceptual meaning about the attitudes, beliefs, or emotions of other people?
- There is extensive evidence that we can use BOLD activity to classify conceptual knowledge using a variety of **concrete** concepts (e.g. *animals* or *tools*; JV Haxby 2012 Neuroimage).
- Only recently there has been a similar interest in studying the brain representations of **abstract** concepts (M Ghio et al. 2016 Neuroimage; Y Wang et al. 2017 PNAS).
- It is unknown how the brain maps different aspects of social information. We investigated the brain representations of social knowledge associated with two fundamental processes in social cognition: **affect** and **likableness** (NH Anderson 1968 J Pers Soc Psychol).
- This fMRI study investigated the representation of abstract social concepts in the human brain using multivariate pattern analysis (MVPA).

Experiment

- Participants (n = 30; mean age 24.07 ± 3.67 years; 18 females) rated on a scale from 0 to 100 how *likable* and *affective* was each social concept before and after the scanning session.
- During the fMRI session, participants listened to short definitions of social concepts and **mentally simulated** another individual behaving the way described in the definition:



- A scanning session consisted on 8 functional runs and a structural run halfway through the session.



- Each functional run contained 36 trials (one trial per social concept; 9 concepts for each concept class). A trial consisted on the auditory presentation of the definition of the social concept for 3.5 seconds followed by another short period of 3.5 seconds to **mentally simulate the referred behavior**.

Decoding the brain representation of social concepts

1. The intraclass correlation coefficient shows a fair test-retest repeatability for the ratings of *affect* and excellent for the ratings of *likableness.*

ID	Concept	Pre	Post	Dimension
sub-001	Boring	20	10	likableness
sub-001	Good-natured	65	70	likableness
sub-001	Thankful	81	72	likableness
sub-001	Cheerful	95	80	likableness
…	…	…	…	
sub-030	Resentful	50	70	affect
sub-030	Sensible	100	95	affect
sub-030	Honest	50	50	affect
sub-030	Lazy	10	0	affect

ICC_{affect} = 0'47 [0,36-0,60]
$ICC_{likableness}$ = 0'93 [0,89-0,96]

2. Ratings of affect *(red)* of concepts related to affective states were significantly higher than those related to non-affective mental states ($t_{(29)}$ = 8.026, p < 0.001, d = 1.465) and ratings of likableness *(grey)* of socially likable concepts were significantly higher than those of socially unlikable concepts ($t_{(29)}$ = 30.382, p < 0.001, d = 5.547).

3. Subjective ratings suggest that the *likableness* of others' behavior can outweigh *affect* on how we represent social knowledge.

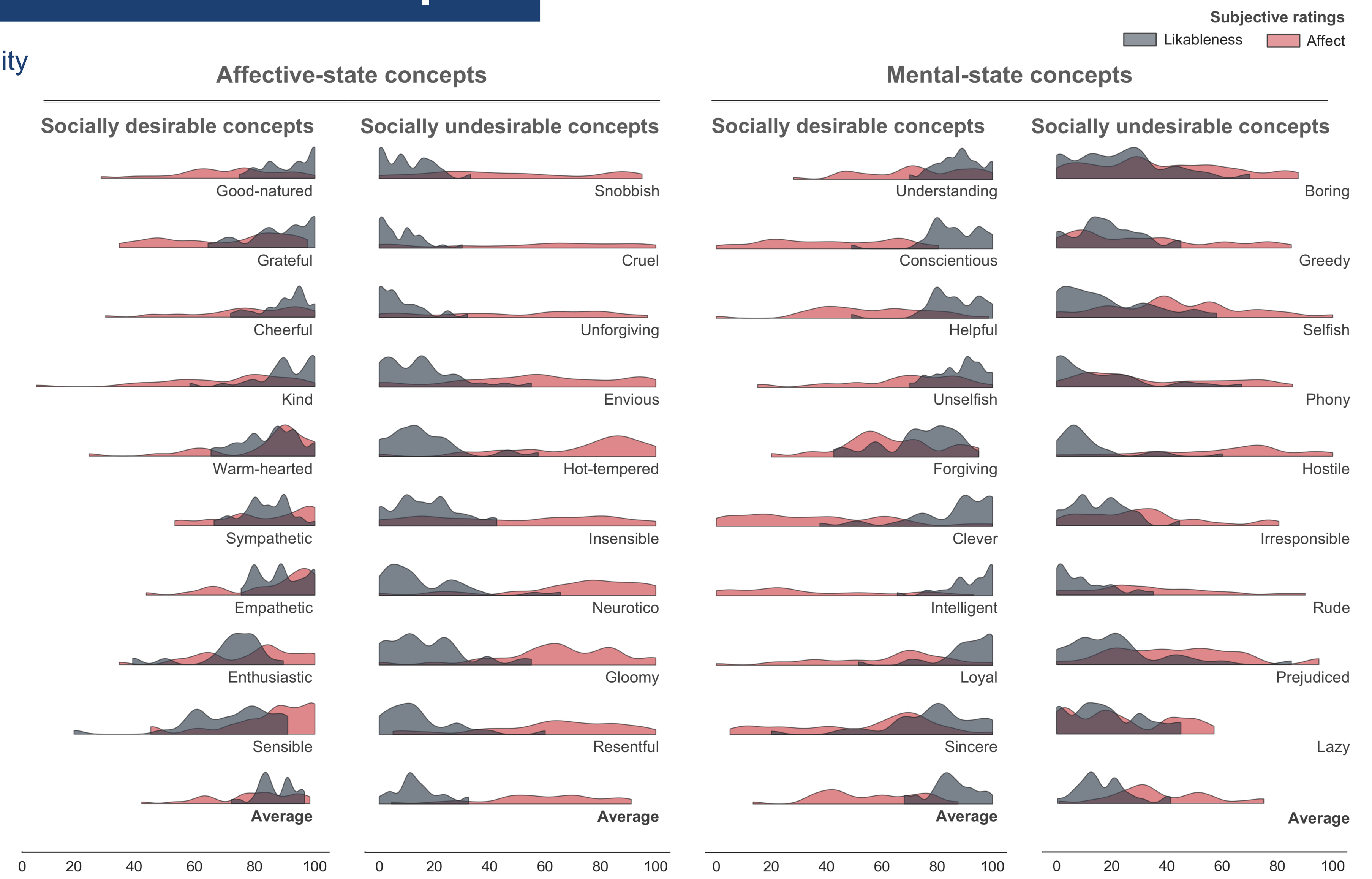


Figure 1. Distribution of ratings of social concepts. Participants read each concept definition and rated whether they described a behaviour involving the emotions of oneself or others (affect; *red*) as well as whether such behaviour was socially likable(*grey*) on a scale from 0 (very nonaffective; very unlikable) to 100 (very affective; very likable).

4. Repeated-measures ANOVAs with one factor (ROI) showed significant differences in decoding accuracy among ROIs for each classification problem *(affect* vs. *likableness*).

		Sum of Squares	df	Mean Square	F	p	ω^2
Partition CV	Affect	0.086	8	0.011	19.505	< .001	0.244
	Likableness	0.240	8	0.030	30.888	< .001	0.267
Item CV	Affect	0.040	8	0.005	8.195	< .001	0.240
	Likableness	0.176	8	0.038	23.531	< .001	0.218

5. A repeated-measures ANOVA with two factors: ROI and dimension of social information (i.e. *affect* vs. *likableness*) showed a significant main effect of dimension for both the cross-validation procedures.

		Sum of Squares	df	Mean Square	F	p	ω^2
Partition CV	ROI	0.289	8	0.036	39.395	< .001	0.314
	Dimension	0.034	1	0.034	7.801	0.009	0.054
	ROI * Dimension	0.037	8	0.005	7.640	< .001	0.056
Item CV	ROI	0.171	8	0.033	27.657	< .001	0.334
	Dimension	0.033	1	0.033	2.938	0.097	0.043
	ROI * Dimension	0.045	8	0.009	7.281	< .001	0.105

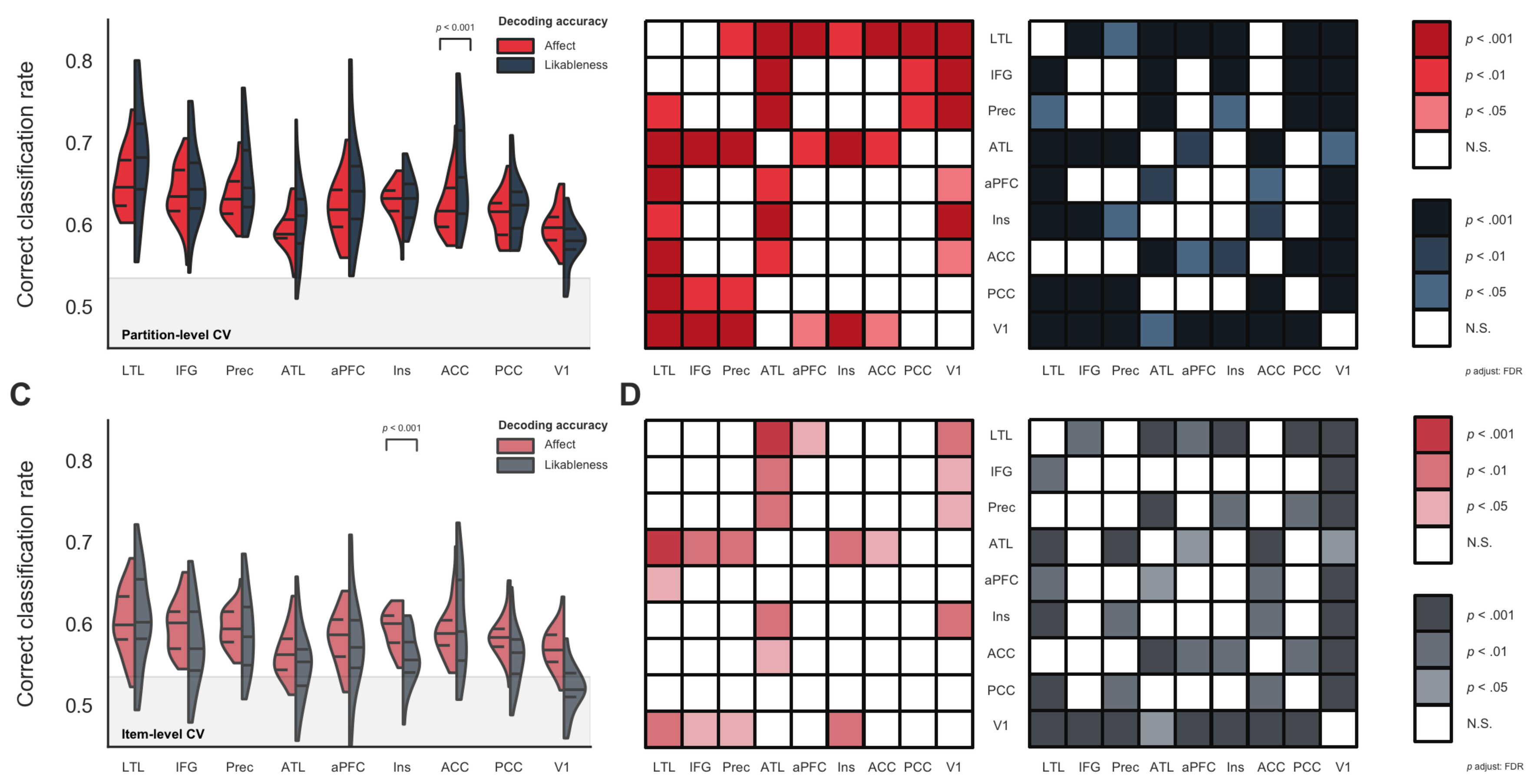


Figure 2. Decoding accuracies of social concepts across the brain. Classification accuracy of both the affect and likableness of social knowledge using partition-level **(A)** as well as item-level **(C)** cross-validation procedures. Panels **(B)** and **(D)** show *post hoc* paired *t*-tests of separate repeated-measures ANOVAs with one factor (ROI) for affect *(red)* and likableness *(grey)* for both CV procedures.

6. *Post hoc* paired *t*-tests showed that the interaction effect was driven by the ACC ($t_{(29)}$ = 4.461, p = 0.001, d = 0.814), with a preference for *likableness* when using the partition-level CV. On the other hand, the effect was driven by the Ins ($t_{(29)}$ = -4.623, p = 0.001, d = 0.844), with a preference for *affect* instead when using the item-level CV.

7. Paired *t*-tests showed that mean classification accuracy was significantly higher in semantic ROIs for both *affect* ($t_{(29)}$ = 5.590, p < .001, d = 1.021) and *likableness* ($t_{(29)}$ = 5.113, p < .001, d = 0.933) using partition-level cross-validation. Similarly, mean decoding accuracy was higher in semantic ROIs for both affect ($t_{(29)}$ = 2.519, p = 0.018, d = 0.460) and likableness ($t_{(29)}$ = 4.133, p < .001, d = 0.755) using item-level cross-validation.

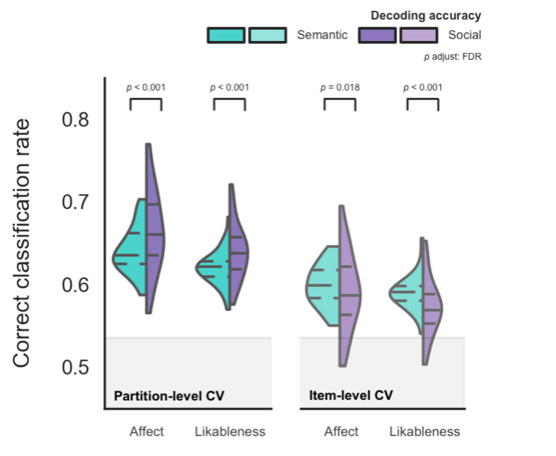
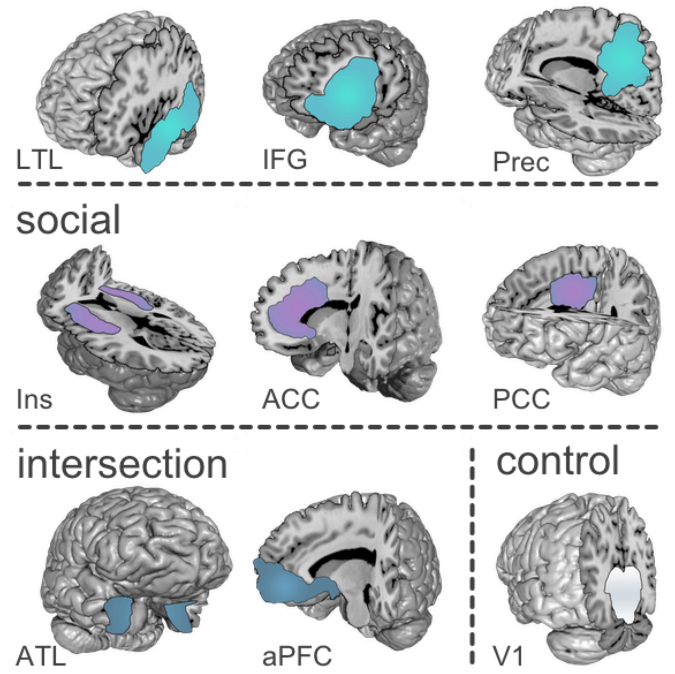


Figure 3. Decoding accuracy of social concepts in semantic vs. social ROIs. Comparison of average decoding accuracy in semantic ROIs with social ROIs using both the partition- and item-level CV procedures. The shaded area indicates the mean empirically estimated chance level (mean = 0.53).

Statistical analysis

- Preprocessing: (i) non-brain tissue removal (BET); (ii) volume realignment (MCFLIRT); (iii) gaussian kernel (FWHM = 3mm) for spatial smoothing; (iv) ICA-based automatic removal of motion artefacts; (v) temporal filtering (high-pass; cutoff = 60s); (vi) coalignment of each session to the 1st session.
- Classification: **SVM-based linear classifier** to decode the brain representation of social knowledge regarding: (i) their *likableness* (high vs. low) and (ii) their *affect* (high vs. low). We used a PCA-based feature selection within each ROI.
- Cross-validation: (i) we used **partitions** of the stacked BOLD data as left-out samples to test the classifier. (ii) we also used entire **items** (i.e. concepts) as left-out sample for testing to better ensure out-of-sample generalization.
- Chance level: we trained a classifier on samples with randomly shuffled labels and tested on samples labelled appropriately to **empirically** estimate chance level performance and used paired *t*-tests to assess statistical significance.
- Statistical significance: decoding performance at the group level was estimated with (i) two repeated-measures ANOVAs with one factor (ROI) to compare ROIs; (ii) with a repeated-measures ANOVA with two factors (ROI and *affect* vs. *likableness*) to analyse whether some ROIs were biased towards decoding the affect or likableness; and (iii) paired *t*-tests to compare the average decoding accuracy in semantic ROIs compared with social ROIs.
- Regions of interest: based on previous studies on semantic (JR Binder et al. 2009 Cereb Cortex) and social information processing (D Alcalá-López et al. 2017 Cereb Cortex):
  + 3 semantic regions (lateral temporal lobe, LTL; inferior frontal gyrus, IFG; and precuneus, Prec)
  + 3 social regions (insula, Ins; anterior cingulate cortex, ACC; posterior cingulate cortex, PCC)
  + 2 semantic & social regions (anterior temporal lobe, ATL; anterior prefrontal cortex, aPFC)
  + 1 control region (primary visual cortex, V1)



Take-home messages

- While previous evidence have analysed temporal correlations between the time series of different ROIs, here we used a pattern classification approach showing these distributed ROIs actually contain information relevant to distinct aspects of social knowledge, beyond just showing activation related to the processing of social information.
- Our results don't support a modular view of the representation of social concepts. Rather, they are consistent with the idea that socially relevant knowledge relies on a widely distributed brain network.

Alcalá-López, D., Smallwood, J., Jefferies, E., Van Overwalle, F., Vogeley, K., Mars, R. B., ... & Bzdok, D. (2017). Computing the social brain connectome across systems and states. *Cerebral cortex*, 28(7), 2207-2232. | **Anderson**, N. H. (1968). Likableness ratings of 555 personality-trait words. *Journal of personality and social psychology*, 9(3), 272. | **Binder**, J. R., Desai, R. H., Graves, W. W., & Conant, L. L. (2009). Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. *Cerebral Cortex*, 19(12), 2767-2796. | **Haxby**, J. V. (2012). Multivariate pattern analysis of fMRI: the early beginnings. *Neuroimage*, 62(2), 852-855. | **Ghio**, M., Vaghi, M. M. S., Perani, D., & Tettamanti, M. (2016). Decoding the neural representation of fine-grained conceptual categories. *Neuroimage*, 132, 93-103. | **Wang**, Y., Collins, J. A., Koski, J., Nugiel, T., Metoki, A., & Olson, I. R. (2017). Dynamic neural architecture for social knowledge retrieval. *Proceedings of the National Academy of Sciences*, 114(16), E3305-E3314.