Spotlight

CellPress

The medial prefrontal cortex in constructing personality models

Jonathan B. Freeman and Ryan M. Stolier

Department of Psychology, New York University, 6 Washington Place, New York, NY 10003, USA

A recent study by Hassabis *et al.* suggests that the brain constructs 'personality models' of other people. When imagining another individual, multi-voxel patterns of fMRI activation in the medial prefrontal cortex (mPFC) contained information about the individual's unique combination of personality traits. The authors propose that, in concert with other regions, the mPFC assembles a model of another's personality that is ultimately used to predict behavior.

Humans are able to quickly acquire complex information about other people and use this information to understand and subsequently predict their behavior. Although the neural mechanisms involved in this process have received considerable attention, neuroimaging research has predominantly focused on how people infer others' transient mental states such as beliefs and intentions. This work has afforded impressive knowledge about the brain basis of 'mentalizing', or theory of mind, including the consistent finding that a network of regions – the mentalizing network – is recruited to represent the transient contents of another's mind [1].

Comparatively few studies have examined how the brain represents another person's dispositional characteristics such as personality traits. Some research suggests that within the mentalizing network, the temporoparietal junction is more involved in the processing of another's transient beliefs, whereas the mPFC integrates such information into higher-order representations such as an individual's enduring traits [2]. A recent study by Hassabis *et al.* [3] took an innovative approach to investigating the brain's representation of personality using neural decoding. The authors propose that the brain constructs 'personality models', or precise representations of an individual's personality that are ultimately used to predict others' behavior.

In the study, participants learned the personalities of four target individuals via a series of statements describing their behavior. The four targets independently differed in the degree (low versus high) to which they embodied two 'Big Five' personality dimensions: agreeableness and extraversion. During a subsequent fMRI session, participants were prompted with a variety of social situations (e.g., spilling one's drink in a bar) and were tasked with imagining how the situations would play out for a given target. Multivariate classification results revealed that neural patterns in the anterior mPFC could reliably discriminate which of the four target individuals a participant was imagining at that time. Specifically, this region appeared to track unique combinations of a target's agreeableness (low or high) and extraversion (low or high). Independently, agreeableness was discriminated in patterns of the posterior cingulate cortex (pCC), while extraversion was discriminated in patterns of the left lateral temporal cortex (LTC) and dorsal mPFC. Connectivity analyses indicated that the anterior mPFC region was functionally coupled with regions discriminating the specific traits. On the basis of these findings, the authors argue that the anterior mPFC constructs a personality model of a unique target individual, while the pCC, left LTC, and dorsal mPFC are involved in processing a target's specific traits. They suggest that the connectivity results demonstrate how the trait information is integrated into a personality model generated by the anterior mPFC.

The notion that the brain may implicitly construct a model of another individual's personality – specifically a model of another's 'Big Five' dimensions – from learned behaviors and access that model while imagining a target individual is a fascinating premise. Previous neuroimaging studies have shown that the mPFC is involved in processing traits implied by behavioral statements (as in the classic 'spontaneous trait inference' paradigm [4]) and in representing a possible 'trait code' [5], but such studies have focused on the mPFC's role in extracting single traits during encoding of behavioral statements. Hassabis *et al.* [3] put forth the striking argument that any number of a target individual's traits may be assembled into a multidimensional personality model integrated in the anterior mPFC and that this model is spontaneously accessed when simulating that individual in a novel situation.

Hassabis *et al.*'s [3] approach and results are exciting, but an important question for future research to address will be what exactly is coded by voxel patterns in their paradigm. For example, if a region's voxel patterns discriminate between thinking of an extraverted versus introverted individual, does this reflect a genuine coding of that underlying trait dimension or some cognitive or evaluative state in response to extraverted versus introverted individuals? The fact that different regions were involved in discriminating extraversion and agreeableness might support the concern that these regions' voxel patterns reflect other mental states or representations that co-vary with thinking about introverted versus extroverted or low- versus high-agreeable targets, respectively. For instance,

Corresponding author: Freeman, J.B. (jon.freeman@nyu.edu).
Keywords: person perception; social neuroscience; mPFC; personality; traits; impressions.

1364-6613/
 http://dx.doi.org/10.1016/j.tics.2014.09.009

such patterns could be involved in discriminating the perceived mood of introverted versus extroverted individuals or an evaluative response to low- versus high-agreeable individuals. Given the results of prior fMRI adaptation studies, one might expect a genuine trait code to be represented in a single region (e.g., mPFC) for multiple traits [5]. The fact that anterior mPFC patterns reflected the unique combination of two dimensions could also potentially reflect covarying states or representations (e.g., different evaluations of an extroverted low-agreeable versus introverted low-agreeable individual) rather than a genuine trait code. Future studies could adopt Hassabis *et al.*'s innovative approach using a larger trait space (e.g., all 'Big Five' dimensions) and using graded rather than discrete levels on these dimensions to more comprehensively probe the nature of personality representation in the mPFC. Such work could also examine multi-voxel patterns associated with traits during on-line encoding (rather than off-line simulation) for a better understanding of the component processes underlying such personality representation.

Undoubtedly, given the authors' argument that the mPFC constructs personality models that are accessed while simulating others' behavior, the stage is now set for future research. Multivariate fMRI in tandem with advanced data-driven techniques [6] could be of great utility in this endeavor, as it may prove difficult to determine if the brain would represent another's personality in the form of the full five-factor model (or other models from behavioral research). For example, prominent models of face evaluation and stereotyping using data-driven techniques have found a different set of dimensions to underlie trait judgments [7,8], and such techniques are ripe for application in understanding neural representations of personality traits. It will be critical to address whether a local unified personality model even exists in the brain, rather than conceptual and trait information retrieved in response to specific individuals (somewhat akin to prototype vs. exemplar models of categorization [9]). Indeed, prior behavioral studies suggest the existence of only limited personality models based on idiosyncratically central traits [10], but Hassabis *et al.*'s novel approach now opens the door to investigating whether a far more comprehensive multidimensional 'personality space' is implicitly constructed by the brain. If true, it would be remarkable.

Acknowledgments

This work was supported in part by a National Science Foundation research grant (NSF-BCS-1423708) awarded to J.B.F.

References

1 Frith, C.D. and Frith, U. (2006) The neural basis of mentalizing. *Neuron* 50, 531–534
2 Van Overwalle, F. (2009) Social cognition and the brain: a meta-analysis. *Brain Hum. Mapp.* 30, 829–858
3 Hassabis, D. *et al.* (2014) Imagine all the people: how the brain creates and uses personality models to predict behavior. *Cereb. Cortex* 24, 1979–1987
4 Uleman, J.S. *et al.* (1996) People as flexible interpreters: evidence and issues from spontaneous trait inference. In *Advances in Social Psychology* (Zanna, M.P., ed.), pp. 211–279, Academic Press
5 Ma, N. *et al.* (2014) Traits are represented in the medial prefrontal cortex: an fMRI adaptation study. *Soc. Cogn. Affect. Neurosci.* 9, 1185–1192
6 Haxby, J.V. *et al.* (2014) Decoding neural representational spaces using multivariate pattern analysis. *Annu. Rev. Neurosci.* 37, 435–456
7 Fiske, S.T. *et al.* (2007) Universal dimensions of social cognition: warmth and competence. *Trends Cogn. Sci.* 11, 77–83
8 Oosterhof, N.N. and Todorov, A. (2008) The functional basis of face evaluation. *Proc. Natl. Acad. Sci. U.S.A.* 105, 11087–11092
9 Smith, E.R. and Zarate, M.A. (1990) Exemplar and prototype use in social categorization. *Soc. Cogn.* 8, 243–262
10 Park, B. *et al.* (1994) Aggregating social behavior into person models: perceiver-induced consistency. *J. Pers. Soc. Psychol.* 66, 437–459