Aligning Brains and Minds

Frank Tong[1,*]
[1]Psychology Department and Vanderbilt Vision Research Center, Vanderbilt University, Nashville, TN 37240, USA
*Correspondence: frank.tong@vanderbilt.edu
DOI 10.1016/j.neuron.2011.10.005

In this issue of *Neuron*, Haxby and colleagues describe a new method for aligning functional brain activity patterns across participants. Their study demonstrates that objects are similarly represented across different brains, allowing for reliable classification of one person's brain activity based on another's.

"How do I know if you see red the same way that I see red? What if you saw all red things the way I see green, but just call those items red?" Even children in primary school seem to appreciate this rather weighty philosophical question, first posed by John Locke (1689). From this simple thought experiment, one could argue that it is impossible to know if the fundamental experiences of one person are truly shared by another. In essence, how can we ever know if our brains or minds are aligned with those around us? Remarkably, advances in human neuroimaging and multivariate pattern analysis could be bringing us a step closer toward addressing questions of this nature.

From a neuroscientific standpoint, it would be difficult to prove that the neural representation of a specific item in one person's brain precisely matched that of another. A more viable approach would be to ask whether the neural representations of many items share a similar functional organization across different brains (e.g., Kriegeskorte et al., 2008). Specifically, one could test whether items that are represented in a more similar manner in one brain are also represented more similarly in another person's brain.

In this issue of *Neuron*, Haxby and colleagues (2011) provide compelling new evidence to suggest that human brains share a very similar representational structure for objects in the world. The authors demonstrate that knowledge of how one person's brain responds to a set of items can greatly facilitate the ability to predict how another person's brain will likely respond to those items. In fact, once a participant's brain activity patterns were brought into functional alignment with the activity patterns of a group template, it was possible to predict what novel object that participant was viewing based on how brains in the reference group responded to those objects. This feat could only be achievable if different brains share similar neural representational structures. How did the authors realize these findings?

An important starting point was to characterize the brain's response to a wide variety of stimuli, to avoid limiting the range of neural representations that might be probed. The authors presented a gripping feature-length movie to participants, *Raiders of the Lost Ark*, because of the rich information contained in such movies and previous work showing that movies evoke similar spatiotemporal patterns of activity across individuals (Hasson et al., 2004). By presenting the same movie to each participant, the resulting brain activity patterns could be used to characterize the shared functional organization across participants. Admittedly, any brief scene in the movie would likely contain multiple stimuli, such as the setting of a cave, a protagonist with a whip, a golden idol resting on an altar, perhaps even a large rolling boulder approaching. Despite the complexity of the stimuli on the screen, each specific time point in the movie could serve as a common index by which to align brain activity patterns across individuals. An implicit assumption to this approach is that activity patterns evoked by multiple stimuli should nonetheless prove effective for characterizing how the brain will likely respond to single objects, new combinations of objects (Macevoy and Epstein, 2009), or even novel objects as long as they share some semantic resemblance to previously viewed stimuli (Mitchell et al., 2008; Naselaris et al., 2009).

Next, the authors had to devise a flexible approach for aligning the brain activity patterns of one individual to another. In most neuroimaging studies, individual brains are aligned to a standard volumetric template, such as the classic Talairach atlas (Talairach and Tournoux, 1988), but such methods fail to account for the fact that the precise location of gyri and sulci can vary considerably from person to person. The gray matter of cortex can be better aligned across subjects by using computational methods to stretch and warp local patches of the cortical surface until the sulci and gyri are well aligned. However, even after cortical alignment, functional brain areas can still vary in size, shape, and location across individuals (Sabuncu et al., 2010). Moreover, functional imaging studies have shown that pattern information can be found at fine spatial scales (Swisher et al., 2010), and such fine-scale information would likely be lost due to imperfect anatomical alignment.

To circumvent the challenges posed by anatomical alignment, the authors developed an entirely different approach of aligning the patterns of functional activity across different brains, a method they call *hyperalignment*. They focused on the ventral temporal cortex, which has been shown to provide detailed information about visual object categories (Haxby et al., 2001). Of critical relevance, the activity patterns in this cortical region convey information primarily about the semantic categories of visual objects rather than their low-level visual properties (Kriegeskorte et al., 2008; Naselaris et al., 2009).

The authors selected 1,000 voxels from the ventral temporal cortex of each participant; among this set of voxels, they could observe distinct spatial patterns of activity for each of the 2000^{+} time points of fMRI data collected during the movie. These spatial patterns of activity can be

analyzed by plotting the response of each voxel along a separate orthogonal dimension, so that any activity pattern can be represented by a single point in this 1,000-dimensional space. Pattern classification methods, such as multivariate pattern analysis (MVPA), can be used to predict what stimulus a person is looking at, given that repeated presentations of a stimulus will evoke very similar patterns of activity within that person's brain. However, a limitation of current MVPA methods is that they usually make far less accurate predictions when applied across individuals, because anatomical coregistration fails to adequately align the functional representations between different brains. What alternatives might there be to devise a mapping between the 1,000-dimensional voxel space of one participant and that of another if anatomy is not taken into account?

Haxby and colleagues (2011) used a specialized algorithm (a *Procrustean transformation*) to rotate and reflect the 1,000-dimensional space of one participant into alignment with that of another, essentially by aligning voxels or combinations of voxels that shared similar time signatures. For example, a voxel that prefers vehicles should respond strongly whenever a car, boat, or airplane appears during the movie; voxels that prefer a different stimulus, such as snakes, should lead to a different time signature in all participants. Because this method involves a rigid transformation of each person's activity patterns, all internal relationships between the similarity of one activity pattern to another will remain preserved. As a consequence, if these activity patterns reflected a higher-order semantic structure shared across participants, this preserved structure might allow for reliable between-subject classification of novel, semantically related stimuli. One by one, the activity patterns of multiple participants were brought into alignment, so that the activity patterns of any new participant could be compared to average functional patterns observed in a large reference group.

How precise was the alignment? To evaluate this, the authors first used activity patterns from one half of the movie to align an individual brain to the reference group, and then attempted to predict what movie segment that person was viewing in the second half of the movie, based on the similarity between that individual's activity pattern and the group's brain responses to the second half of the movie. The level of between-subject classification was very high, reaching ~70% accuracy where chance-level performance would have been less than one percent. The authors further found that they could reduce the dimensionality of the group activity patterns to 35 distinct principal components and still achieve excellent classification performance. This implies that 30 or so dimensions were sufficient to capture the range of information contained in these brain responses to the movie.

Hyperalignment based on the movie data also allowed for successful classification of novel static objects presented in a separate experiment. In one experiment, between-subject classification was used to differentiate human faces, monkey faces and dog faces. In another experiment, the authors used between-subject classification to discriminate between six animal species (ladybug beetles, luna moths, mallard ducks, yellowthroated warblers, ring-tailed lemurs, and squirrel monkeys). Strikingly, the accuracy of between-subject classification proved to be as good as within-subject classification, that is, training and testing a pattern classifier on a participant's own brain activity. The fact that it was possible to generalize to novel objects based on other people's brain data suggests that the ventral temporal cortex represents objects in a similar manner across individuals. When errors in classification did occur, they often occurred among semantically similar items, such as ducks and warblers, and appeared equally prevalent for within- and between-subject classification.

Although previous studies have demonstrated that brain activity patterns reflect the semantic similarity of objects, the present study goes further to show that this semantic organization is broadly similar across individuals. It would be intriguing to extend this work in a variety of directions. For example, current fMRI models that predict an individual's brain responses to novel words or scenes (Mitchell et al., 2008; Naselaris et al., 2009) could be extended to investigate whether such stimuli are represented similarly across participants. Hyperalignment might also be used to ask how similar one person's neural representations are to those of others. For example, there is some evidence to suggest that the degree of correlated activity found between a speaker (telling a story) and a listener depends on how well the listener understood the story (Stephens et al., 2010). Perhaps hyperalignment could be used to enhance studies of the neural bases of story comprehension and human communication. It has also been reported that individuals with autism exhibit more idiosyncratic patterns of brain activity in response to movies (Hasson et al., 2009). Hyperalignment might be used to test whether these differences are attributable to differential attention or eye movements or to genuine differences in the underlying meaning of objects to these individuals. Finally, it would be worth testing whether hyperalignment based on one type of movie would prove effective for between-subject classification of a movie that differs greatly in style and image content, such as a nature documentary. A recent study demonstrated remarkably accurate predictions of how the early visual cortex of individual participants would respond to novel movies, based on how these visual areas responded to the local motion signals contained in a variety of movie clips (Nishimoto et al., 2011). This vision-based approach to analyzing brain activity, although highly powerful, should be considered quite distinct and complementary to the semantics-based approach emphasized by the present study.

To revisit John Locke's armchair experiment, if he were here today, would he find these neuroimaging results convincing in their suggestion that people represent the world in a very similar way? Based on the knowledge of his time, Locke was careful to argue that color experiences might be reversed across individuals according to an inverted spectrum, so that the similarity relationships between any two colors (and the ease with which they could be discriminated) should remain the same. We now know that the human eye registers color information through three different color-sensitive photoreceptors, and these signals are further recombined to form red-green and blue-yellow

opponent color mechanisms. Behavioral testing could therefore be used to tell apart whether a person perceived colors according to a normal or inverted spectrum. However, it would be difficult or impossible to tell if someone experienced a reversal along a color-specific axis, such as red and green (Palmer, 1999). In the present study, Haxby and colleagues (2011) found that 30^{+} dimensions were needed to attain high accuracy of object predictions across participants. It remains a logical possibility that any one of those dimensions might have been precisely reversed in one of the participants tested. For example, inanimate objects might have evoked a greater feeling of "life" than animate objects in an idiosyncratic participant. However, if the mathematical fitting of highly complex, multidimensional data worked extremely well across individuals, most scientists would consider the possibility of such a perfectly reversed mapping to be implausible. A more reasonable conclusion would be that similar representational structures exist in the brains, and minds, of different individuals. Indeed, John Locke himself concluded that despite the logical possibility of a reversal of experiences, "I am nevertheless very apt to think that the sensible ideas produced by any object in different men's minds, are most commonly very near and undiscernibly alike" (Locke, 1689).

REFERENCES

Hasson, U., Nir, Y., Levy, I., Fuhrmann, G., and Malach, R. (2004). Science *303*, 1634–1640.

Hasson, U., Avidan, G., Gelbard, H., Vallines, I., Harel, M., Minshew, N., and Behrmann, M. (2009). Autism Res. *2*, 220–231.

Haxby, J.V., Gobbini, M.I., Furey, M.L., Ishai, A., Schouten, J.L., and Pietrini, P. (2001). Science *293*, 2425–2430.

Haxby, J.V., Guntupalli, J.S., Connolly, A.C., Halchenko, Y.O., Conroy, B.R., Gobbini, M.I., Hanke, M., and Ramadge, P.J. (2011). Neuron *72*, this issue, 404–416.

Kriegeskorte, N., Mur, M., Ruff, D.A., Kiani, R., Bodurka, J., Esteky, H., Tanaka, K., and Bandettini, P.A. (2008). Neuron *60*, 1126–1141.

Locke, J. (1689). An Essay Concerning Human Understanding (London: William Tegg).

Macevoy, S.P., and Epstein, R.A. (2009). Curr. Biol. *19*, 943–947.

Mitchell, T.M., Shinkareva, S.V., Carlson, A., Chang, K.M., Malave, V.L., Mason, R.A., and Just, M.A. (2008). Science *320*, 1191–1195.

Naselaris, T., Prenger, R.J., Kay, K.N., Oliver, M., and Gallant, J.L. (2009). Neuron *63*, 902–915.

Nishimoto, S., Vu, A.T., Naselaris, T., Benjamini, Y., Yu, B., and Gallant, J.L. (2011). Curr. Biol. *21*, 1641–1646.

Palmer, S.E. (1999). Behav. Brain Sci. *22*, 923–943, discussion 944–989.

Sabuncu, M.R., Singer, B.D., Conroy, B., Bryan, R.E., Ramadge, P.J., and Haxby, J.V. (2010). Cereb. Cortex *20*, 130–140.

Stephens, G.J., Silbert, L.J., and Hasson, U. (2010). Proc. Natl. Acad. Sci. USA *107*, 14425–14430.

Swisher, J.D., Gatenby, J.C., Gore, J.C., Wolfe, B.A., Moon, C.H., Kim, S.G., and Tong, F. (2010). J. Neurosci. *30*, 325–330.

Talairach, J., and Tournoux, P. (1988). Co-planar Stereotaxic Atlas of the Human Brain (New York: Thieme).