PHILOSOPHICAL TRANSACTIONS B

rstb.royalsocietypublishing.org

Review

Cite this article: Adolphs R, Nummenmaa L, Todorov A, Haxby JV. 2016 Data-driven approaches in the investigation of social perception. *Phil. Trans. R. Soc. B* **371**: 20150367.
http://dx.doi.org/10.1098/rstb.2015.0367

Accepted: 25 January 2016

One contribution of 15 to a theme issue 'Attending to and neglecting people'.

Subject Areas:
cognition

Keywords:
social neuroscience, social perception, ecological validity, intersubject brain correlation, face space

Author for correspondence:
Ralph Adolphs
e-mail: radolphs@hss.caltech.edu

Data-driven approaches in the investigation of social perception

Ralph Adolphs[1], Lauri Nummenmaa[2,3], Alexander Todorov[4] and James V. Haxby[5]

[1]Division of the Humanities and Social Sciences, California Institute of Technology, Pasadena, CA, USA
[2]Department of Neuroscience and Biomedical Engineering, School of Science, Aalto University, Espoo, Finland
[3]Turku PET Centre and Department of Psychology, University of Turku, Turku, Finland
[4]Department of Psychology, Princeton University, Princeton, NJ, USA
[5]Department of Psychological and Brain Sciences, Dartmouth College, Hanover, NH, USA

The complexity of social perception poses a challenge to traditional approaches to understand its psychological and neurobiological underpinnings. Data-driven methods are particularly well suited to tackling the often high-dimensional nature of stimulus spaces and of neural representations that characterize social perception. Such methods are more exploratory, capitalize on rich and large data-sets, and attempt to discover patterns often without strict hypothesis testing. We present four case studies here: behavioural studies on face judgements, two neuroimaging studies of movies, and eyetracking studies in autism. We conclude with suggestions for particular topics that seem ripe for data-driven approaches, as well as caveats and limitations.

1. Introduction

Psychological and neuroscience studies on social cognition have traditionally aimed at hypothesis-driven studies and rigorous control by using well-defined stimuli that are amenable to clear experimental manipulation. Despite its undisputed success, this approach comes with several inherent limitations. First, typical experimental stimuli are simply not as good as natural stimuli at eliciting reliable, robust responses. As the brain has been tuned to respond to a continuous sensory stream during evolution, complex natural stimuli trigger more reliable neural responses than the conventionally used, well-controlled yet simplified stimuli. For example, cells in the cat visual cortex respond more strongly to natural pictures than to random patterns [1]. Similarly, movies but not noise patches or sinusoidal gratings enhance response reliability in visual cortex [2]. In humans, natural, dynamic faces also activate the face-processing network more consistently than static or rigidly moving faces [3,4]. More practically, participants in fMRI experiments are more motivated to participate, pay better attention, and can tolerate longer experimental sessions when they are watching a naturalistic and engaging movie than when looking at visual gratings.

Second, because the typical stimuli used in experiments are not representative of the real world, there is *prima facie* doubt about whether findings from them would generalize to the real world. This worry is indeed valid: responses to complex stimuli cannot necessarily be predicted from straightforward combinations of responses to simple stimuli [5]. The responses may also be categorically different: many psychological phenomena simply cannot be made to fit into fully controlled, traditional stimulus models. For example, listening to a musical piece [6], perception of complex action sequences [7], and social interaction [8] all span several overlapping and hierarchical time scales involving parallel processing of overlapping sensory features. Consequently, they cannot be adequately explored with classic experimental designs that use decompositions of these stimuli. Similarly, there is the historical effort to find 'simpler' explanations of what features drive the responses of neurons or brain regions, with selectivity for high-dimensional stimuli like faces. Just trying to cut up a face into parts does not result in an explanation of the

THE ROYAL SOCIETY PUBLISHING

responses to whole faces, and instead many stimuli seem to be processed in a way that appears 'holistic'.

But holistic responses do not emerge out of nowhere, so there must still be an underlying mechanism that explains how they are generated from more basic features or dimensions of some sort. Thus, when researchers generate stimulation models and design experimental conditions, they may not necessarily know what are the most important stimulus dimensions that should be manipulated. This is where data-driven approaches can really show their power: they can be used for revealing what defines an 'optimal stimulus', and what defines basic features or dimensions that may synthesize such a stimulus. Because the stimulus space is typically too high-dimensional, and the underlying dimensions may be too abstract for us to come up with good *a priori* hypotheses about them, the mechanisms behind responses to complex stimuli may not be amenable to *a priori* theorizing in an efficient way. Yet data-driven approaches are unconstrained by such hypotheses and instead can show us stimuli, features, and dimensions that we would never have thought of. One of the simplest examples of how this might work conceptually is reverse correlation. In this approach, the stimuli are unstructured noise. On each trial, the noise stimulus will, just by chance, generate some similarity to features of the target stimulus. When a face-selective neuron is shown thousands of such noise masks and we keep track of where exactly all the pixels in the noise image are on each trial, we can simply do a spike-triggered average that shows the association between the response of the neuron with the locations of all the pixels in the images. That can subsequently be used, for instance, to generate a composite image that begins to resemble a face. It is important to note here that even very exploratory, data-driven approaches usually incorporate some kind of regularization—we assume linearity or Gaussian distributions, or make other assumptions about the form the stimuli might take, or the mapping from stimuli to behaviour. While data-driven methods are exploratory in the sense that they eschew a rigid hypothesis about the conclusions of the analyses, they still rely on varied background assumptions that are often important to note.

A conceptually similar approach to reverse correlation can be taken when the dependent measure is behavioural classification and the stimulus space pertains to a particular category, rather than to unstructured noise. In this kind of experiment, subjects are shown a large number of stimuli and asked to sort them into categories—say different emotions that faces express. The stimuli in this case are drawn from a structured stimulus space and are noisy or sparsely sampled. Once again, over many trials, the statistical relationship between random sources of variation in the stimuli and the behavioural classification is extracted [9]. The sources of variation can take a number of different forms: one could simply inject noise into an underlying base stimulus to ask how a noisy feature can be interpreted in different ways. For example, Mona Lisa's smile may look neutral or happy, depending on slight variations of the sensory input that we can identify if we average over a sufficiently large number of stimuli [10]. One could also introduce random sampling of the spatial location, or the spatial frequency, or the time of occurrence of the stimulus and its features [9]. In all these cases, we are being unbiased in not providing a specific hypothesis to begin with, and by introducing a source of random variation (but keeping track of that variation, trial by trial) we are letting the data tell us about the relationship between stimulus variation and response category. One need not use noise as in these examples, but can instead use any rich, high-dimensional stimulus that does not set up a categorical hypothesis to begin with.

This kind of data-driven approach is equally useful when it comes to making sense of complex responses to our complex stimuli. The challenge often amounts to one of dimensionality reduction, based on all the data available, a big issue especially for EEG, MEG, and fMRI data that measure brain responses. Behavioural responses can also be high-dimensional, but we in effect have the subject do the dimensionality reduction by specifying a specific mapping: push one of two buttons, etc. As with completely data-driven approaches on the stimulus end, the difficulty is interpreting the relevance of the component dimensions that are found: what, psychologically, do these mean?

The solution is of course to find an approach to link stimuli to behaviour (or brain response). The stimuli and the fMRI responses are individually high-dimensional, and for each we have dimensionality reduction methods available. But we want to mesh the two: we want to find those dimensions that actually matter to behaviour or to psychology. There are a host of methods available to do this, and the examples below give an overview of how this can work.

2. Case study 1: studies of face space to discover the dimensions that underlie human social judgements

In the first systematic study of social judgements from faces in modern social psychology, Secord *et al.* [11] concluded that 'the conventional 'elementalizing' used by psychologists in seeking to explain their data is simply inappropriate for physiognomy, and that new ways of conceptualizing physiognomy need to be found if it is to be fully understood.' The researchers reached this conclusion after finding out that the same facial feature in the context of other facial features can acquire completely different meaning and, consequently, influence judgements in opposite ways. Lips with the same thickness, for example, in one combination create the expression of meekness and in another of excitability and conceitedness. These effects have been subsequently described as illustrating the holistic perception of faces [12–14].

The standard experimental approach is to create all possible feature combinations and to test how these combinations influence judgements. However, the space of possible combinations is intractably large. With more than two features, the possible combinations rapidly proliferate. With just 10 binary features, we have 1024 feature combinations. With 20 binary features, we have 1 048 576 feature combinations. To make things worse, we do not even know what constitutes a proper facial feature. Our intuitions point to things like eyes and mouth but each of these 'features' could be further broken down into a number of smaller features such as pupil size, sclera size, sclera coloration, thickness of lips, shape and bushiness of eyebrows, and so on. And the features are not binary. As a result of this complexity, typical experiments focus on a set of features with limited variation.

In contrast with the standard approach, the data-driven approach does not manipulate features and need not

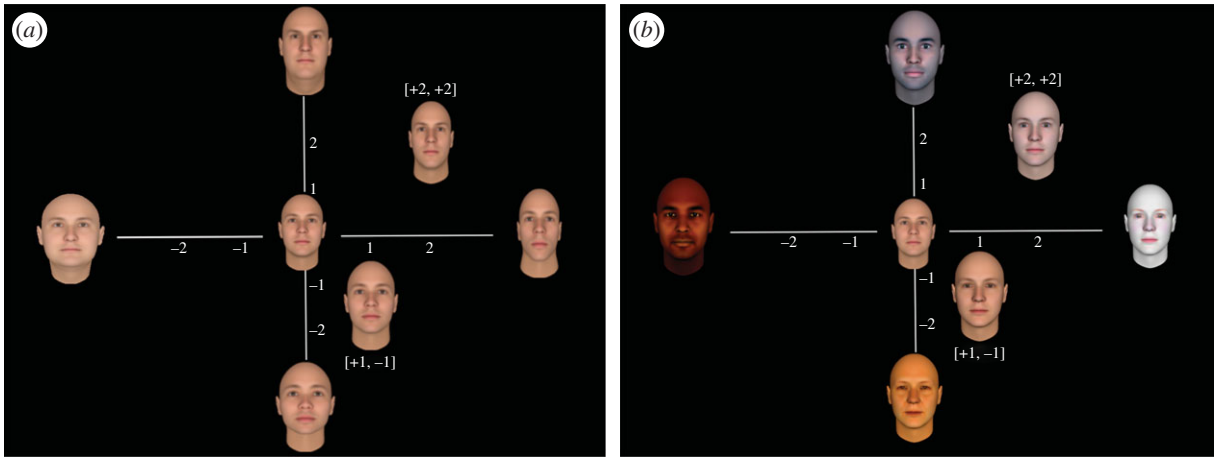

Figure 1. Statistical models of faces. An individual face stimulus can be represented as a vector in a dimensional space. With synthetic face stimuli, one can omit dimensions that would be psychologically irrelevant, such as the type of camera taking the picture, and incorporate a more restricted set of dimensions to begin with. (*a*) Illustration of statistical face space with two dimensions representing face shape. (*b*) Illustration of statistical face space with two dimensions representing face reflectance.

constrain the search for combinations of features to subsets of features. The starting point for the data-driven approach is a mathematical model of face representation that captures the holistic variation of faces. The first such models were based on a principal components analysis (PCA) of the pixel intensities of face images [15]. Subsequent models were based on a PCA of the shape of faces acquired from three-dimensional laser scanning of faces [16,17]. The same technique can be used to build a statistical model of face reflectance, texture, and pigmentation, using the red, green, and blue colour values from each pixel on the face's surface. In these models, each face is represented as a vector in a multi-dimensional space (figure 1). The statistical face space allows us to randomly sample faces that are representative of the face variation captured by these models. To identify the combinations of features that lead to social judgements, we simply need to ask participants to judge the randomly sampled faces. If these judgements are reliable, we can build a model of the face variation that drives the judgements [18–20]. This model is a new vector in the statistical face space that captures the meaningful face variation with respect to the judgement.

Figure 2 shows such a model based on subject's ratings of faces. In the case of trustworthiness judgements, the primary dimension on which faces are evaluated [18,21], we can see that trustworthy faces are more feminine and have positive expressions. Note that although emotional expressions were not manipulated, they naturally emerged from the judgements of the randomly varying faces. We can infer that weak emotional signals are an important input to judgements, and test this in standard experiments [22]. In the case of dominance judgements, the second fundamental dimension on which faces are evaluated, we can see that dominant faces are more masculine and facially mature. We can infer that inferences of physical strength are an important input to dominance judgements [18,21].

Coming back to the example quote we gave at the beginning, there are clear future directions here: notably, extending the analysis to nonlinear effects. The results from linear techniques, such as the PCA approach we reviewed above, can be taken as initial findings that could guide more complex studies that begin to explore truly holistic face processing in which different features or dimensions interact in more complicated ways. Needless to say, this opens up a much larger search space, but if suitably guided by initial findings, particular parts of this space could be explored. This last

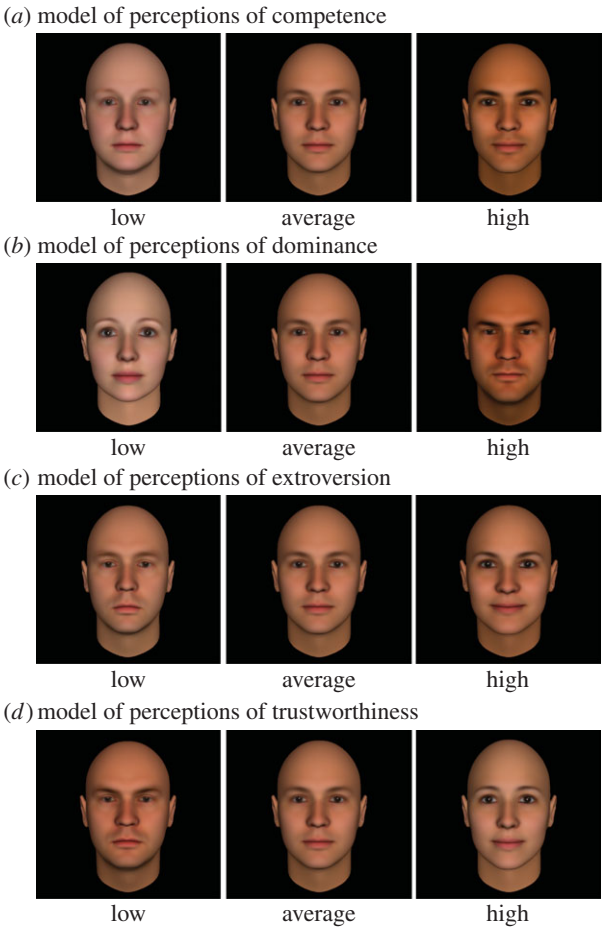

Figure 2. Faces generated by a data-driven computational model of judgements of (*a*) competence; (*b*) dominance; (*c*) extroversion; and (*d*) trustworthiness. The middlemost face on each row is the average face in the statistical model. The face to the right is 3SD above the average face on the respective trait dimension; the face to the left is 3SD below the average face.

point also raises an important comment on the relationship between data-driven and hypothesis-driven methods: they should interact and inform one another. Not all methods should be entirely agnostic, but as we accrue findings, future approaches, even if data-driven in part, should incorporate aspects of prior findings to help constrain our search for those features and dimensions that matter the most.

3. Case study 2: neural representations of movies

The above example shows how one can map even complex stimuli-like faces into a space, relate dimensions of that space to social judgements, and use this information to construct new synthetic stimuli. However, even these stimuli are still impoverished compared with the real world. What happens if we try to take account of a dynamic, multi-modal stimulus?

Cinema provides an excellent way for simulating real life in the laboratory. Movies are multi-modal, engaging snapshots of reality, often describing human interactions in realistic conditions. From the researcher's point of view, however, this realism comes at a cost: when multiple overlapping stimulus features are present simultaneously, stimulus-specificity of the corresponding neural responses is difficult to disentangle. This is even more of a problem when we try to use a technique such as fMRI, which tracks sluggish hemodynamic responses as the basic dependent measure, even though it has the great advantage of a whole-brain field of view. Yet, the multidimensionality of the stimulus and the long, variable haemodynamic time courses of fMRI can be exploited with novel data analytic techniques, which have emerged in parallel with increases in computational power available for analyses.

Early fMRI studies confirmed that modelling blood-oxygen-level dependent (BOLD) data with subject classification of events, such as occurrence of faces and voices in movies, results in clear patterns of functionally specialized neural responses [23,24]. Specific subsystems—such as those involved in face or voice perception—thus operate during natural viewing as in more controlled experiments (see also [25] for similar results in macaques). On the fMRI end, it is also possible to use the pattern of activation across multiple voxels as the response, rather than the more typical mass-univariate response, which averages over voxels. This multivariate pattern analysis can be used to classify stimuli using tools from machine learning, and it can be used to visualize stimulus categories in terms of the pattern similarity of the response evoked. The multivariate approach to classification was first introduced by Haxby in 2001 [26] and has now become nearly ubiquitous in fMRI studies. The approach of visualizing similarity spaces, pioneered by Nikolaus Kriegeskorte, so-called representational similarity analysis [27,28], is also becoming a mainstream tool for discovering structure in high-dimensional stimulus spaces on the basis of the similarity of the neural activation response that they evoke.

The most recent work combines many of the above aspects and has revealed that voxel-wise cortical representations of thousands of different, overlapping categories can be resolved from fMRI data acquired during movie viewing [29], a feat that would require literally days of scanning using conventional designs with a separate experimental condition for each tested category. However, interpretation of the observed activation patterns in relation to the overlapping, time-locked stimulus features remains challenging [29,30] (see the final section).

(a) Stimulus-blind analysis with intersubject correlations

Prolonged naturalistic stimuli such as movies provide an additional window to human brain function by enabling analysis of intersubject reliability or intersubject synchronization of the brain activity time courses. This involves extracting voxelwise time courses from each participant, and averaging the voxelwise correlation of time courses across each possible subject pair. Temporal accuracy of the signal is thus sacrificed for the sake of gaining sufficient signal-to-noise ratio for quantifying regional response reliability across subjects. Critically, such analysis does not assume anything regarding the underlying sensory features of the stimulus; thus it can be used for exploring the regional response properties in the brain. These kind of fMRI studies have revealed that human cortical activity is time-locked across individuals (at the time-scale of a few seconds) during naturalistic audiovisual stimulation, confirming that neural processes occur at similar temporal time scales across individuals while processing naturalistic events presented in videos [31,32] or in spoken narratives [33,34]; recently, it has also been applied with significantly higher frequencies in MEG [35]. Intersubject similarity measures can also be extended with reverse-correlation techniques for probing the functional organization of the human brain. Instead of using a pre-specified stimulation model, it is possible to extract haemodynamic time series from a specific brain region, and go back to the original stimulus to assess whether focal, consistent brain signals are associated with specific stimulus features [31]. Such explorative approaches open up insights into the organization of human brain function that would go unnoticed with *a priori* stimulation models.

Response reliability can also be quantified within subjects for repeated presentations of the same stimulus. Such work has revealed two broad, distinct sets of brain networks—one whose responses are consistent and a second whose responses are inconsistent with external stimulation, thus probably reflecting 'intrinsic' and 'extrinsic' modes of information processing [36]. In line with this, early visual areas show reliable responses independently of disruption of temporal structure of events in movies, whereas disruption significantly lowers synchronization of upstream areas such as posterior superior temporal sulcus and frontal eye fields [37]. This suggests that different cortical systems integrate sensory information at different time scales. These findings are also supported by frequency-specific intersubject-correlation analyses of movie viewing data, which find that sensory cortices show synchronization at fastest, and frontal cortices at slowest frequencies [38].

Finally, intersubject-correlation analysis can also be extended to allow computation of moment-to-moment time series of intersubject similarity, which can be used to model how similarity in neural activation across participants fluctuates as a function of time or due to experimental conditions [39]. Conceptually, regionally selective synchronization of brain activity across individuals could be the elementary mechanism supporting mutual understanding of the social environment. Activity within individual people's brains indeed becomes increasingly synchronous in a regionally selective fashion when they feel similar, strong emotions [40] or assume similar psychological perspectives towards the events described in a movie [41,42].

(b) Independent components analysis

An alternative solution for parsing the associations between brain activity and overlapping stimulus dimensions in movies involves the use of independent component analysis (ICA, figure 3, [43]). In this blind signal separation approach,

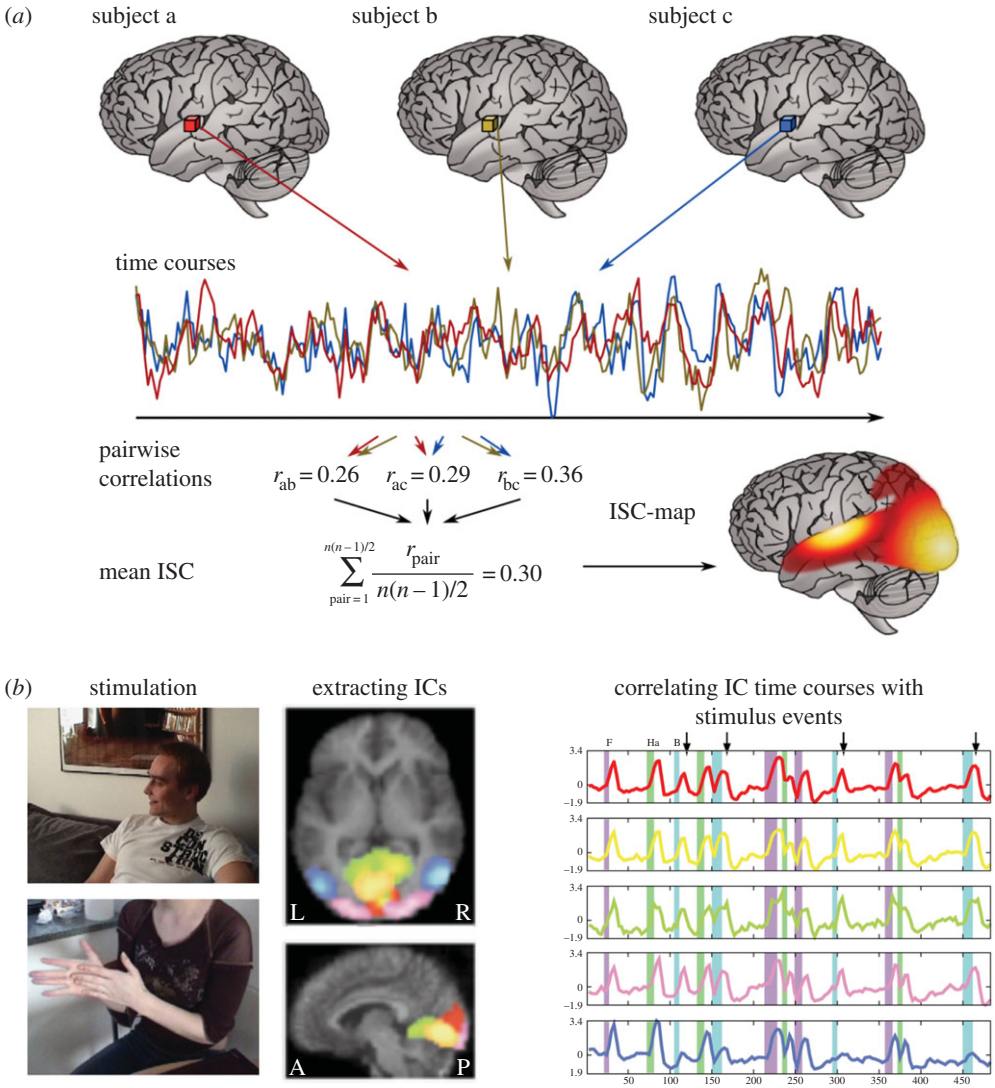

Figure 3. (*a*) Stimulus-blind analysis using intersubject correlation (ISC) is based on temporal similarity of voxelwise time courses across subjects. When computed in a sliding window across the time course it also allows linking moment-to-moment ISC with a stimulus model for quantifying the relationship between external stimulus and response reliability across subjects. (*b*) Independent components analysis is based on dividing the BOLD signal into statistically independent components. As in ISC, the extracted components can subsequently be linked with stimulus events. (*a*) Courtesy of Juha Lahnakoski, (*b*) adapted with permission from Malinen *et al*. [32].

the BOLD signal is treated as a mixed signal, which is mathematically divided into statistically independent signals, that is, independent components (ICs). The brain voxels belonging to a single component thus share temporal activation patterns, and areas with strong anatomical connections (e.g. language areas) also often show associated response patterns [44]. ICA consequently allows anatomically unconstrained, functional parcellation of the brain activity patterns, and enables revealing brain activation patterns that may have been impossible to predict beforehand. It should be noted that ICA does not inherently provide dimensionality reduction (although dimensionality is typically first reduced for such analyses, using methods such as PCA), but rather does so through a process of *post hoc* selection of those components from the ICA that are deemed relevant data (for instance, components due to 'noise' such as head movement artefacts are removed using ICA). Yet after the functional subdivision, the ICs can be regressed against any stimulation model to reveal their functional significance [32]. This approach may also resolve the complex relationship between combinations of overlapping stimulus features and resultant brain activation patterns: after extracting functional patterns (ICs), temporal dependencies between stimulus features and brain activation patterns can be established using canonical correlation analysis [45].

ICA is particularly powerful when analysing brain activation patterns associated with complex high-dimensional stimuli, where onsets and offsets of discrete stimulus evens cannot be defined. For example, tasks such as viewing a movie, navigation in a VR environment, or simulated driving involve adaptation to a constantly changing environment with multiple perceptual, attentional, and motor behaviours, yet the task structure cannot be fully specified *a priori*, practically precluding all sorts of model-based analyses of the brain imaging data (see review in [46]). Despite its promise, ICA retains a number of problems in application and interpretation in the context of neuroimaging. First, determining the number of components as well as aligning components across subjects is far from straightforward. Second, interpretation of the components still often requires a formal stimulation model, so that IC time courses can be linked with, for example, stimulation events.

4. Case study 3: using movie data to build a model of shared representational spaces

As we discussed in the previous section, movies provide a much richer and more diverse sampling of stimuli than is used in more controlled experiments. They, therefore, also provide a better basis for modelling the structure of high-dimensional representational spaces that is shared across brains [47,48], thus allowing us to align different people's brain representations in a functional space that circumvents problems with standard anatomical alignment. A model of the commonality of the functional architecture of the human brain that is based on structural anatomical features is inadequate. Anatomy-based alignment does not align the fine-scale patterns of activity, embedded in fine-scale features of cortical topographies, which carry distinctions between the information represented by those patterns [47,49–51]. Moreover, anatomy-based alignment does not capture the idiosyncratic individual variability of coarse topographic features, such as the location, size and conformation of the borders of functional areas such as retinotopically organized early visual areas, motion-sensitive MT or category-selective areas in ventral temporal cortex.

A common model of the functional architecture that captures these features—fine-scale patterns of activity and individual variability of coarse-scale features—has been developed using a new algorithm, hyperalignment, and achieves broad general validity by estimating model parameters based on responses to a complex, dynamic, naturalistic stimulus, such as a full-length movie [47,51,52]. The elements of this model are a common, high-dimensional representational space and individual transformation matrices that project data from idiosyncratic, individual anatomic spaces into the common model space. The first demonstration of this model was a common representational space for ventral temporal cortex [47]. Subsequent developments have extended this approach to make a common model of representational spaces in all of human cortex [51]. The original algorithm used the Procrustes transformation [53] to derive the individual transformation matrices and common model space. More recently, a probabilistic algorithm appears to afford building a model with even better performance [52].

The common model finds response-tuning profiles that are shared across brains. Between-subject correlations of local time-series responses to movies double after hyperalignment [47,51]. Between-subject multivariate pattern classification (bsMVPC) is dramatically higher after hyperalignment, as compared to bsMVPC of anatomically aligned data, equalling, and at times exceeding, within-subject MVPC. This counterintuitive result is achievable because the common model affords larger, multi-subject datasets for training pattern classifiers, whereas training datasets for wsMVPC are necessarily limited to data from single subjects.

Using responses to a naturalistic movie to derive the common model space and estimate the parameters for individual-specific hyperalignment transformation matrices greatly increases the general validity of the common model in two ways. First, surprisingly, the responses to the movie are more distinctive than are responses to more controlled stimuli, such as isolated, single objects or faces on a blank background. In a matched bsMVPC analysis of single time points from responses to movies as compared to responses to isolated faces and objects, accuracies were over twice as high for responses to the movie than for responses to the more controlled stimuli (figure 4, adapted from [47]). Second, responses to the movie afford a common model with far greater general validity than is possible if the common model is derived based on responses to a more controlled and limited range of stimuli. The hyperalignment algorithm can also be applied to data obtained while subjects engage in controlled experiments. Whereas a model based on movie data generalizes to afford high levels of bsMVPC for unrelated experiments, a model based on data from a controlled experiment does not (figure 5, adapted from [47]).

Shared representation is the basis of social communication. Modelling how different brains represent the same information, therefore, is key for understanding the epistemological basis of social cognition. The use of complex, dynamic, naturalistic stimuli provides a stronger basis for modelling that common basis than does the use of more controlled stimuli. This advantage is due to several factors. First, rich, naturalistic movies sample a much broader range of stimuli and include dynamic movement, language, social interactions, and narrative structure. This broader sampling affords more precise estimation of parameters for a high-dimensional model of shared representation that has broad-based validity across experiments. Moreover, this sampling includes high-level aspects of social cognition that play a limited role in controlled experiments. Second, viewing a continuous movie provides strong predictions at each moment about what to expect next at all levels of representation, from low-level visual features of changes due to continuous movement, to high-level semantic features based on plot and character development. This allows predictive coding to operate in a natural way, and prediction signals may be a key component of neural representation. By contrast, controlled experiments are generally designed to render prediction signals as uninformative as possible, making them inconsistent and a probable source of noise.

5. Case study 4: modelling attention in autism

For our last example, we return once more to analysis of stimuli, this time with complex images. As mentioned at the beginning, one thing we would like to know when faced with a complex stimulus, is the contribution made by each of its constituents. This was studied in an eye-tracking experiment, which asked what the influence is that different features make to people's fixations, and in particular, how this might differ in people with autism spectrum disorder. Conceptually, this is like a big regression problem: we want to predict where people fixate on an image, and we want to have as regressors all the different factors that could influence this. Those factors include some basic low-level or 'bottom-up' properties of images, such as the centre location of the image, regions that have high contrast or colour, and so forth. They also include object- and semantic-level attributes of the images, such as the locations of round or square objects, the identity of those objects (faces, text, cars, etc.), and other judged properties (e.g. their emotional relevance). The low-level or pixel-level attributes can be assigned using computer algorithms mimicking the response properties of the visual system from retinal neurons through V1 [54], whereas the object-level and semantic-level attributes currently require manually annotating this on each image.

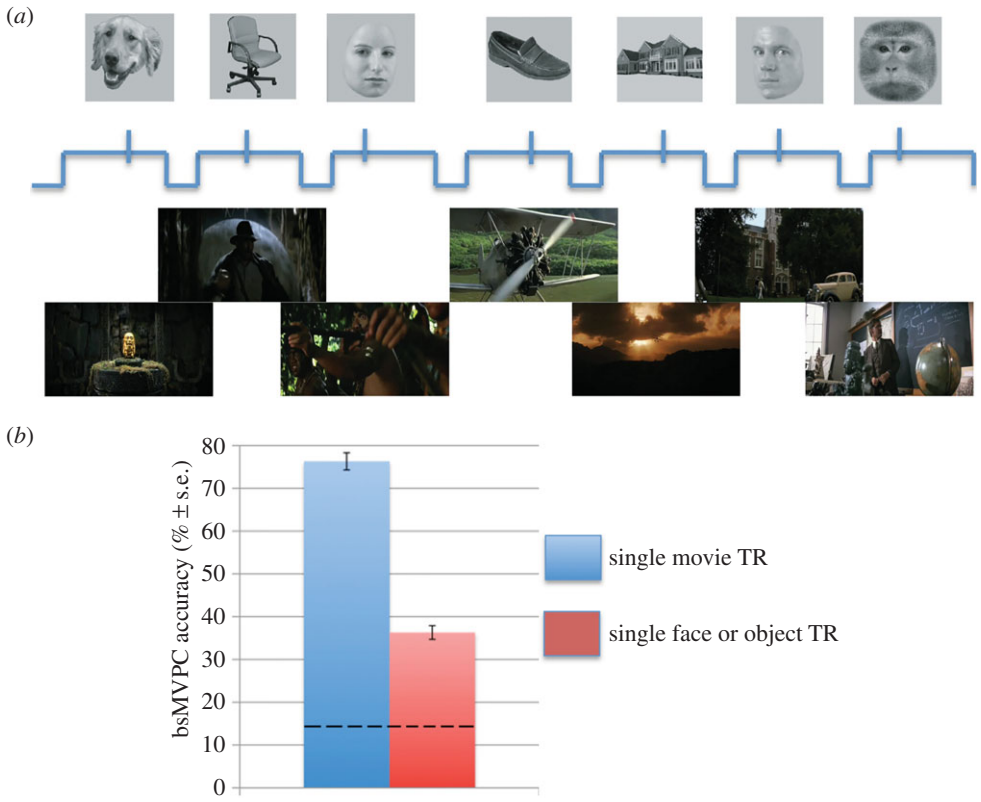

Figure 4. bsMVPC of single time points (TRs) from a movie, *Raiders of the Lost Ark*, and from a face and object perception study. (*a*) A seven-way bsMVPC analysis matched the probability of correct classifications for the two experiments. Response patterns from one TR in each stimulus block of a run in face and object experiment were extracted from all subjects. This was repeated for all eight runs. Response patterns of TRs during the movie presentation at the same acquisition time as selected for the face and object experiment were extracted from all subjects to perform a similar seven-way bsMVPC analysis. (*b*) Results showed that BSC accuracy for movie time points was more than twice that for time points in the face and object perception experiment. Dashed lines indicate chance classification (one out of seven). Adapted from Haxby *et al*. [47].

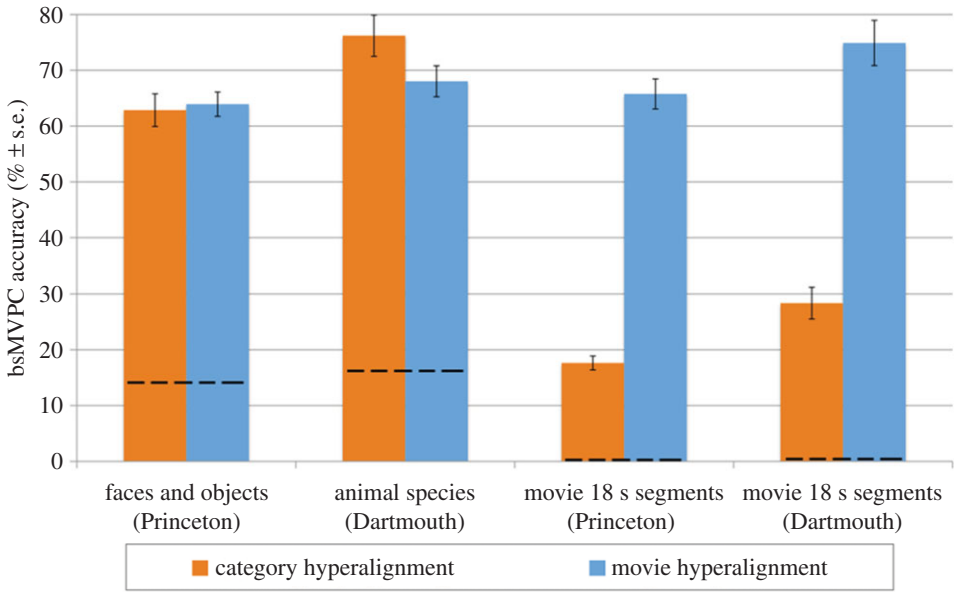

Figure 5. Comparison of the general validity of common models based on responses to the movie and on responses to still images. Common models were built based on responses to a movie, *Raiders of the Lost Ark*, and responses to single images in a face and object category perception experiment [47], performed at Princeton, and an animal species perception experiment (Connolly *et al*. [49]), performed at Dartmouth. Results on the left show bsMVPC accuracies for the responses to single faces, objects, and animal species. Results on the right show bsMVPC accuracies for 18 s time segments in the movie. Note that common models based on responses to the category images afford good bsMVPC for those experiments but do not generalize to bsMVPC of responses to movie time segments. Only the common model based on movie viewing generalizes to high levels of bsMVPC for stimuli from all three experiments. Dashed lines indicate chance performance. From Haxby *et al*. [47].

In the experiment, we used an out-of-sample prediction, a support vector machine classification that was trained, in each subject, on a subset of the eye-tracking data and then evaluated on the remainder of the fixations [55]. Subjects viewed 700 images that had over 5000 objects annotated on them. The analysis yields 'weights' that, when normalized, correspond to how much influence each feature has on where a subject fixates. We found some expected weights in the control subjects: for instance, faces had a high weight, as did text, as these are features that typically strongly attract attention. However, for the autism group, there were systematic differences in how bottom-up saliency versus scene semantics guided eye movements: Notably, they remained, across fixations, more influenced by low-level properties (such as brightness), and less influenced by object- and semantic-level properties (such as faces; [56]).

These findings illustrate the power of using data-driven techniques also to investigate psychiatric disorders. There is acknowledgement that the current diagnostic way of classifying psychiatric disorders probably needs revision and should be based on more objective criteria. Data-driven approaches to analysing behaviour, eye tracking, or brain responses could provide such criteria, perhaps also allowing us to view psychiatric disorders more dimensionally.

6. Limitations

There are a number of caveats to data-driven approaches for the investigation of social perception. In many cases—such as in reverse-correlation techniques—large numbers of trials are required, and in all cases, rich stimuli and/or data are required. This imposes constraints on the time required for an experiment, the number of sensory channels engaged, and the number of dependent measures obtained. In the case of using fMRI while subjects are watching a movie, the process can be rendered efficient because there is a large bandwidth both at the stimulus end (complex, dynamic, audiovisual stimulus) and the dependent measure of brain activity (parallel measures in 50 000 voxels all over the brain). In the case of more standard reverse-correlation approaches, the approach can, however, quickly become very inefficient and require a huge number of trials. Relatedly, statistical power, reliability, and replication are important considerations. In general, one would like to set up an analysis that includes out-of-sample prediction (such as cross-validation), and there are now many approaches from machine learning to apply this to fMRI and other data.

Some kind of regularization or dimensionality reduction is often needed at the front end to begin any analysis, because the number of datapoints is typically much larger than the number of exemplars in a category. For instance, the number of subjects or of stimuli may be quite small in relation to the richness of measures obtained. A number of such methods, such as ICA that was discussed above, are available for this. It is also often advantageous to have an independent set of data from which a more guided analysis can be conducted. For instance, in fMRI studies, it is often informative to have a separate session that yields specific regions of interest or data on which alignment is based.

A major conceptual limitation with all data-driven methods is that, by design, they eschew a prior hypothesis or theory. As such, it becomes a challenge to know what meaning to assign to the results discovered: what do the extracted dimensions mean? Sometimes emerging dimensions or components make intuitive sense or may be directly linked with the applied stimulation model, but sometimes not. For instance, when applying ICA to data, it is often difficult to know which of the components produced reflect artefactual data (e.g. due to movement of the subject in the scanner or activity driven by eye movements) and which are psychologically meaningful data. As the method itself is blind to this question, it is the final interpretation of the experimenter that must decide.

There are several solutions to this situation. One is of course to have some prior knowledge, perhaps results from other convergent experiments, that can help to triangulate on meaning in the results (see e.g. [29] for linking model-free and model-based responses to semantic categories). A second is to use data-driven methods that nonetheless discover functions describing the data and thus give further insight. An example of this latter approach is symbolic regression, which uses genetic algorithms to generate a function that classifies the data. To our knowledge, this approach has, however, never been applied to fMRI data.

To illustrate this problem a little more, let's return to our first case study of the face space. If we just show synthetic faces on a computer screen, we can completely describe the stimuli. In images consisting of 500×500 pixels, the corresponding 500×500 matrix of grayscale values describes the stimulus completely. One might think this is a fairly easy problem: just do a principal component analysis on all the faces, from their initial 500×500 pixelwise representations. Doing so, however, will still give many dimensions with possibly erroneous ordering of how those dimensions contribute to all the variability in the faces. If one makes some faces slightly darker, and some lighter, or shoot some with a Polaroid and others with a digital camera, a completely stimulus-driven approach will pull out dimensions such as the type of camera and the global brightness of the image. Yet those are not likely to be psychologically relevant dimensions. Instead, psychologically relevant dimensions could be: those people that are familiar to me, versus those that are not. How on earth could we find this dimension solely by analysing the pixel-by-pixel features of the stimuli?

This then shows us a big hurdle in applying data-driven methods. We cannot simply operate on the stimuli because there are complex associations that people have with subsets of stimuli, and because certain aspects of stimuli (like the type of camera) are mostly irrelevant to people. That is the problem: we do not know what is relevant in our high-dimensional stimulus set. The brain does. That is why we need to link stimuli with brain or behaviour in some way and, for example, generate novel stimuli using the emerging dimensions, and go back to the good old-fashioned practice of asking subjects to group stimuli and label the groups, or simply evaluate what the stimuli look like.

Authors' contributions. All authors contributed equally.

Competing interests. We have no competing interests.

Funding. L.N. is supported by The Academy of Finland (MIND program grant 265917) and European Research Council (starting grant no. 313000). R.A. is supported by the National Institute of Mental Health (Conte Center grant).

Acknowledgement. We thank Riitta Hari, local students in Finland, and the Attention and Performance series board for helping organize the conference on which this paper is based.

References

1. Touryan J, Felsen G, Dan Y. 2005 Spatial structure of complex cell receptive fields measured with natural images. *Neuron* **45**, 781–791. (doi:10.1016/j.neuron.2005.01.029)
2. Yao HS, Shi L, Han F, Gao HF, Dan Y. 2007 Rapid learning in cortical coding of visual scenes. *Nat. Neurosci.* **10**, 772–778. (doi:10.1038/nn1895)
3. Fox CJ, Iaria G, Barton JJS. 2009 Defining the face processing network: optimization of the functional localizer in fMRI. *Hum. Brain Mapp.* **30**, 1637–1651. (doi:10.1002/hbm.20630)
4. Schultz J, Brockhaus M, Bülthoff HH, Pilz KS. 2013 What the human brain likes about facial motion. *Cereb. Cortex* **23**, 1167–1178. (doi:10.1093/cercor/bhs106)
5. Felsen G, Dan Y. 2005 A natural approach to studying vision. *Nat. Neurosci.* **8**, 1643–1646. (doi:10.1038/nn1608)
6. Alluri V, Toiviainen P, Jääskeläinen IP, Glerean E, Sams M, Brattico E. 2012 Large-scale brain networks emerge from dynamic processing of musical timbre, key and rhythm. *NeuroImage* **59**, 3677–3689. (doi:10.1016/j.neuroimage.2011.11.019)
7. Zacks JM, Speer NK, Swallow KM, Maley CJ. 2010 The brain's cutting-room floor: segmentation of narrative cinema. *Front. Hum. Neurosci.* **4**, 12. (doi:10.3389/fnhum.2010.00168)
8. Stephens GJ, Silbert LJ, Hasson U. 2010 Speaker–listener neural coupling underlies successful communication. *Proc. Natl Acad. Sci. USA* **107**, 14 425–14 430. (doi:10.1073/pnas.1008662107)
9. Schyns PG, Petro LS, Smith ML. 2007 Dynamics of visual information integration in the brain for categorizing facial expressions. *Curr. Biol.* **17**, 1580–1585. (doi:10.1016/j.cub.2007.08.048)
10. Kontsevich LL, Tyler CW. 2004 What makes Mona Lisa smile? *Vis. Res.* **44**, 1493–1498. (doi:10.1016/j.visres.2003.11.027)
11. Secord PF, Dukes WF, Bevan W. 1954 Personalities in faces: I. An experiment in social perceiving. *Genetic Psychol. Monogr.* **49**, 231–279.
12. Rossion B. 2013 The composite face illusion: a whole window into our understanding of holistic face perception. *Vis. Cogn.* **2**, 139–253. (doi:10.1080/13506285.2013.772929)
13. Todorov A, Loehr V, Oosterhof NN. 2010 The obligatory nature of holistic processing of faces in social judgments. *Perception* **39**, 514–532. (doi:10.1068/p6501)
14. Young AW, Hellawell D, Hay DC. 1987 Configurational information in face perception. *Perception* **16**, 747–759. (doi:10.1068/p160747)
15. Turk M, Pentland A. 1991 Eigenfaces for recognition. *J. Cogn. Neurosci.* **3**, 71–86. (doi:10.1162/jocn.1991.3.1.71)
16. Blanz V, Vetter T. 1999 A morphable model for the synthesis of 3D faces. In *Proc. of the 26th Annu. Conf. on Computer Graphics and Interactive Techniques*, pp. 187—194.
17. Blanz V, Vetter T. 2003 Face recognition based on fitting a 3D morphable model. *IEEE Trans. Pattern Anal. Mach. Intell.* **25**, 1063–1074. (doi:10.1109/TPAMI.2003.1227983)
18. Oosterhof NN, Todorov A. 2008 The functional basis of face evaluation. *Proc. Natl Acad. Sci. USA* **105**, 11 087–11 092. (doi:10.1073/pnas.0805664105)
19. Todorov A, Oosterhof NN. 2011 Modeling social perception of faces. *IEEE Signal Process. Mag.* **28**, 117–122.
20. Walker M, Vetter T. 2009 Portraits made to measure: manipulating social judgments about individuals with a statistical face model. *J. Vis.* **9**, 12, 1–13. (doi:10.1167/9.11.12)
21. Todorov A, Said CP, Engell AD, Oosterhof NN. 2008 Understanding evaluation of faces on social dimensions. *Trends Cogn. Sci.* **12**, 455–460. (doi:10.1016/j.tics.2008.10.001)
22. Said C, Sebe N, Todorov A. 2009 Structural resemblance to emotional expressions predicts evaluation of emotionally neutral faces. *Emotion* **9**, 260–264. (doi:10.1037/a0014681)
23. Bartels A, Zeki S. 2004 Functional brain mapping during free viewing of natural scenes. *Hum. Brain Mapp.* **21**, 75–85. (doi:10.1002/hbm.10153)
24. Moran JM, Wig GS, Adams Jr RB, Janata P, Kelley WM. 2004 Neural correlates of humor detection and appreciation. *NeuroImage* **21**, 1055–1060. (doi:10.1016/j.neuroimage.2003.10.017)
25. Russ BE, Leopold DA. 2015 Functional MRI mapping of dynamic visual features during natural viewing in the macaque. *NeuroImage* **109**, 84–94. (doi:10.1016/j.neuroimage.2015.01.012)
26. Haxby JV, Gobbini MI, Furey ML, Ishai A, Schouten JL, Pietrini P. 2001 Distributed and overlapping representation of faces and objects in ventral temporal cortex. *Science* **293**, 2425–2429. (doi:10.1126/science.1063736)
27. Kriegeskorte N, Bendettini P. 2007 Analyzing for information, not activation, to exploit high-resolution fMRI. *NeuroImage* **38**, 649–662. (doi:10.1016/j.neuroimage.2007.02.022)
28. Mur M, Bandettini PA, Kriegeskorte N. 2009 Tools of the trade: revealing representational content with pattern-information fMRI—an introductory guide. *Soc. Cogn. Affect. Neurosci.* **4**, 101–109. (doi:10.1093/scan/nsn044)
29. Huth AG, Nishimoto S, Vu AT, Gallant JL. 2012 A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron* **76**, 1210–1224. (doi:10.1016/j.neuron.2012.10.014)
30. Lahnakoski JM, Glerean E, Salmi J, Jääskeläinen I, Sams M, Hari R, Nummenmaa L. 2012 Naturalistic fMRI mapping reveals superior temporal sulcus as the hub for the distributed brain network for social perception. *Front. Hum. Neurosci.* **6**, 14. (doi:10.3389/fnhum.2012.00233)
31. Hasson U, Nir Y, Levy I, Fuhrmann G, Malach R. 2004 Intersubject synchronization of cortical activity during natural vision. *Science* **303**, 1634–1640. (doi:10.1126/science.1089506)
32. Malinen S, Hlushchuk Y, Hari R. 2007 Towards natural stimulation in MRI—Issues of data analysis. *NeuroImage* **35**, 131–139. (doi:10.1016/j.neuroimage.2006.11.015)
33. Nummenmaa L, Saarimäki H, Glerean E, Gotsopoulos A, Jääskeläinen IP, Hari R, Sams M. 2014 Emotional speech synchronizes brains across listeners and engages large-scale dynamic brain networks. *NeuroImage* **102**, 498–509. (doi:10.1016/j.neuroimage.2014.07.063)
34. Wilson SM, Molnar-Szakacs I, Iacoboni M. 2008 Beyond superior temporal cortex: intersubject correlations in narrative speech comprehension. *Cereb. Cortex* **18**, 230–242. (doi:10.1093/cercor/bhm049)
35. Lankinen K, Saari J, Hari R, Koskinen M. 2014 Intersubject consistency of cortical MEG signals during movie viewing. *NeuroImage* **92**, 217–224. (doi:10.1016/j.neuroimage.2014.02.004)
36. Golland Y, Bentin S, Gelbard H, Benjamini Y, Heller R, Nir Y, Hasson U, Malach R. 2007 Extrinsic and intrinsic systems in the posterior cortex of the human brain revealed during natural sensory stimulation. *Cereb. Cortex* **17**, 766–777. (doi:10.1093/cercor/bhk030)
37. Hasson U, Yang E, Vallines I, Heeger DJ, Rubin N. 2008 A hierarchy of temporal receptive windows in human cortex. *J. Neurosci.* **28**, 2539–2550. (doi:10.1523/JNEUROSCI.5487-07.2008)
38. Kauppi J-P, Jääskeläinen IP, Sams M, Tohka J. 2010 Inter-subject correlation of brain hemodynamic responses during watching a movie: localization in space and frequency. *Front. Neuroinform.* **4**, 12.
39. Glerean E, Salmi J, Lahnakoski JM, Jaaskelainen IP, Sams M. 2012 Functional magnetic resonance imaging phase synchronization as a measure of dynamic functional connectivity. *Brain Connect.* **2**, 91–101. (doi:10.1089/brain.2011.0068)
40. Nummenmaa L, Glerean E, Viinikainen M, Jaaskelainen IP, Hari R, Sams M. 2012 Emotions promote social interaction by synchronizing brain activity across individuals. *Proc. Natl Acad. Sci. USA* **109**, 9599–9604. (doi:10.1073/pnas.1206095109)
41. Lahnakoski JM, Glerean E, Jääskeläinen IP, Hyönä J, Hari R, Sams M, Nummenmaa L. 2014 Synchronous brain activity across individuals underlies shared psychological perspectives. *NeuroImage* **100**, 316–324. (doi:10.1016/j.neuroimage.2014.06.022)
42. Nummenmaa L, Smirnov D, Lahnakoski JM, Glerean E, Jaaskelainen IP, Sams M, Hari R. 2014 Mental action simulation synchronizes action-observation circuits across individuals. *J. Neurosci.* **34**, 748–757. (doi:10.1523/JNEUROSCI.0352-13.2014)
43. Hyvarinen A, Oja E. 2000 Independent component analysis: algorithms and applications. *Neural Netw.* **13**, 411–430. (doi:10.1016/S0893-6080(00)00026-5)
44. Bartels A, Zeki S. 2005 Brain dynamics during natural viewing conditions—a new guide for

mapping connectivity *in vivo*. *NeuroImage* **24**, 339–349. (doi:10.1016/j.neuroimage.2004.08.044)

45. Ylipaavalniemi J, Savia E, Malinen S, Hari R, Vigário R, Kaski S. 2009 Dependencies between stimuli and spatially independent fMRI sources: towards brain correlates of natural stimuli. *NeuroImage* **48**, 176–185. (doi:10.1016/j.neuroimage.2009.03.056)
46. Calhoun VD, Pearlson GD. 2012 A selective review of simulated driving studies: combining naturalistic and hybrid paradigms, analysis approaches, and future directions. *NeuroImage* **59**, 25–35. (doi:10.1016/j.neuroimage.2011.06.037)
47. Haxby JV, Guntupalli JS, Connolly AC, Halchenko YO, Conroy BR, Gobbini MI, Hanke M, Ramadge PJ. 2011 A common, high-dimensional model of the representational space in human ventral temporal cortex. *Neuron* **72**, 404–416. (doi:10.1016/j.neuron.2011.08.026)
48. Haxby JV, Connolly AC, Guntupalli JS. 2014 Decoding neural representational spaces using multivariate pattern analysis. *Annu. Rev. Neurosci.* **37**, 435–456. (doi:10.1146/annurev-neuro-062012-170325)
49. Connolly AC, Guntupalli JS, Gors J, Hanke M, Halchenko YO, Wu Y-C, Abdi H, Haxby JV. 2012 The representation of biological classes in the human brain. *J. Neurosci.* **32**, 2608–2618. (doi:10.1523/JNEUROSCI.5547-11.2012)
50. Cox DD, Savoy RL. 2003 Functional magnetic resonance imaging (fMRI) "brain reading": detecting and classifying distributed patterns of fMRI activity in human visual cortex. *NeuroImage* **19**, 261–270. (doi:10.1016/S1053-8119(03)00049-1)
51. Guntupalli JS, Hanke M, Halchenko YO, Connolly AC, Ramadge PJ, Haxby JV. In press. A model of representational spaces in human cortex. *Cereb. Cortex*.
52. Chen P-H, Chen J, Yeshurun-Dishon Y, Hasson U, Haxby JV, Ramadge PJ. 2015 A reduced-dimension fMRI shared response model. In *The Annu. Conf. on Neural Information Processing Systems* (*NIPS*).
53. Schönemann P. 1966 A generalized solution of the orthogonal procrustes problem. *Psychometrika* **31**, 1–10. (doi:10.1007/BF02289451)
54. Itti L, Koch C. 1998 A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* **20**, 1254–1259. (doi:10.1109/34.730558)
55. Xu J, Jiang M, Wang S, Kankanhalli MS, Zhao Q. 2014 Predicting human gaze beyond pixels. *J. Vis.* **14**, 28. (doi:10.1167/14.1.28)
56. Wang S, Jiang M, Duchesne XM, Laugeson EA, Kennedy DP, Adolphs R, Zhao Q. 2015 Atypical visual saliency in autism spectrum disorder quantified through model-based eye tracking. *Neuron* **88**, 604–616. (doi:10.1016/j.neuron.2015.09.042)