Revealing the multidimensional mental representations of natural objects underlying human similarity judgements

Martin N. Hebart, Charles Y. Zheng, Francisco Pereira and Chris I. Baker

Objects can be characterized according to a vast number of possible criteria (such as animacy, shape, colour and function), but some dimensions are more useful than others for making sense of the objects around us. To identify these core dimensions of object representations, we developed a data-driven computational model of similarity judgements for real-world images of 1,854 objects. The model captured most explainable variance in similarity judgements and produced 49 highly reproducible and meaningful object dimensions that reflect various conceptual and perceptual properties of those objects. These dimensions predicted external categorization behaviour and reflected typicality judgements of those categories. Furthermore, humans can accurately rate objects along these dimensions, highlighting their interpretability and opening up a way to generate similarity estimates from object dimensions alone. Collectively, these results demonstrate that human similarity judgements can be captured by a fairly low-dimensional, interpretable embedding that generalizes to external behaviour.

We live in a world full of objects that we can identify, place into different categories, communicate and reason about, and act on in a meaningful manner. These abilities are remarkable, given that our ever-changing environment requires us to constantly map unique sensory information from the things around us to our internal representations of objects, categories and concepts. To carry out this mapping and make sense of our world, we therefore need to determine the similarity between the sensory information emanating from the environment and our internal mental representations. Not surprisingly, similarity has been suggested to play an important role in elucidating the structure of our mental representations and can help explain how we recognize objects, form categories, structure our conceptual knowledge and predict the behaviour of our visual world on the basis of our experience. Moreover, representational similarities offer a useful tool to relate behaviour, computational models and brain activity patterns.

Despite the success of similarity and the wide use of similarity judgements for studying mental representations of objects, similarities alone offer only an indirect and mostly descriptive view of the format of our mental representations. They can inform us about the degree to which two or more representations are similar, but they are agnostic as to what properties (or dimensions) each representation is made up of and what dimensions are shared between those representations. For example, most people would agree that a dog and a cow are more similar than a dog and a car, probably because dogs and cows share more relevant dimensions, such as being animate, natural or soft. To understand the structure of our mental representations of objects, we need to identify those core dimensions that form the basis of our similarity judgements. These dimensions need to fulfil two criteria. First, they should be predictive of behaviour and thus able to characterize the mental representational space. Second, to move beyond description and provide understanding, we need to identify a set of dimensions (from the infinite number possible) that can be interpreted meaningfully.

Here, we present a computational model of mental representations of objects based on a large-scale assessment of human similarity judgements for natural object images. Prior experimental, neuropsychological and neuroimaging evidence have led to the proposal of object dimensions such as animacy, manipulability and real-world size, but these dimensions describe only a selective and largely incomplete portion of our mental representational space. In contrast to traditional small-scale experimental approaches, which often use artificial stimuli or words, we collected a large number of similarity judgements for images of 1,854 different objects, capturing both visual and conceptual mental representations for a wide, representative range of natural objects. Rather than relying on explicit verbal reports of what object features are perceived as relevant, the model learns these dimensions directly from these similarity judgements.

Using this data-driven approach, we identify 49 dimensions underlying similarity judgements that lead to excellent predictions of both single-trial behaviour and similarity scores between pairs of objects. We demonstrate that the dimensions are meaningful and characterize the large-scale structure of our mental representations of objects. The model allows for the accurate prediction of categorization behaviour, while within categories, individual dimensions reflect object typicality. Finally, we demonstrate that human participants can use these dimensions directly to provide good predictions of similarity judgements, underscoring the interpretability of dimensions and offering a first step towards a generative model of the perceived similarity of natural objects.

Results

To characterize the representational space of natural objects, we had to overcome several obstacles. First, we needed to identify a set of objects that is representative of the objects encountered in the real world. For that purpose, we chose the 1,854 objects in the THINGS database, which we developed to provide a comprehensive list of
living and non-living things according to their everyday use in the American English language. For each object, we chose a representative image that had been shown to be named consistently during the creation of this database. The advantage of using images rather than words is that they may provide additional purely perceptual information that is relevant for judging the similarity of objects and that might not come to mind immediately when using words.

Second, we needed to identify a task that would allow us to best quantify the similarity between pairs of objects. Ideally, this task would highlight all relevant dimensions contributing to the similarity of pairs of objects and would be independent of the context in which these objects appear\textsuperscript{12,13}. While pairwise similarity ratings on a Likert scale are one of the most popular approaches, this task implicitly assumes that all dimensions relevant to judging the similarity of pairs of objects are always and immediately available to the observer, even when the objects are very dissimilar and may seem to have nothing in common. Here we chose a different approach, in which we concurrently presented three object images \(i, j\) and \(k\) in a triplet odd-one-out task (Fig. 1a). By choosing the odd-one-out object, participants indicate which pair of objects \((i,j)\), \((i,k)\) or \((j,k)\) is the most similar among this set. The key benefit of this task is that the third object always serves as a context for the other two objects, thus highlighting the relevant dimensions that make two objects most similar. By repeatedly varying the third object for a given pair of objects, we are thereby implicitly sampling across a wide range of contexts in which the objects might be encountered. We can then express similarity as an approximation of the probability \(p(i,j)\) of participants choosing objects \(i\) and \(j\) together, irrespective of context. In addition, since the similarity of objects is determined with respect to all other objects, this approach naturally constrains the number of possible dimensions to those relevant for discriminating among objects\textsuperscript{10}.

Third, we needed to collect sufficient data with these objects and this task. While the odd-one-out task provides a principled approach for investigating similarity across contexts, for 1,854 objects it would require \(\sim1.06\) billion combinations of triplet judgements for a single estimate of the full similarity matrix. This would make conducting the odd-one-out task at this scale not feasible. However, if similarity depends only on a small number of independent dimensions, it should be possible to approximate the entire similarity matrix.
with only a fraction of those judgements. In this study, we sampled 1.46 million unique responses from 5,301 workers using the online platform Amazon Mechanical Turk (Fig. 1a), pooling all responses across workers (median number of responses per worker, 60). This corresponded to 0.14% of possible unique trials.

Our goal was to build a computational model that is capable of predicting behaviour in the odd-one-out task, that captures the similarity between all pairs of objects and that provides interpretable object dimensions. At the centre of this model is a representational embedding, which is a quantitative characterization of objects as vectors in a multidimensional representational space. This embedding can be described as a matrix X, in which each column corresponds to a dimension and each row to an object vector across all dimensions (Fig. 1b). In the context of our model, this embedding should allow us to (1) predict behaviour for individual trials not included in the training data and (2) generate the entire similarity matrix between all pairs of objects.

To create this embedding, we made two key assumptions. First, we assumed that dimensions are sparse, which is a reasonable assumption, given that not all dimensions are expressed in all objects. For example, for a putative dimension of animacy, a cardboard box would probably have a value of 0. Second, we assumed that dimensions are continuous and positive. Accordingly, the numeric value of an object for a given dimension could be interpreted as the degree to which the dimension is expressed in the object, which should support interpretability.

The modelling procedure was as follows (Fig. 1c). We initialized the model with 90 random dimensions, assumed that after model fitting, sparsity would reduce the dimensionality of the embedding to a smaller number. For a given triplet (i,j,k) for which we had collected a behavioural judgement, we then calculated the dot product between the embedding vectors of all three pairs of objects (i,j), (i,k) and (j,k). Accordingly, when two objects express high values for many dimensions, this measure yields a large number, whereas when one object expresses high values in dimensions for which the other object expresses low values, this measure yields a small number. Next, on the basis of those three dot products, we estimated the probability of choosing one of the three pairs of objects in this context, which is equivalent to the third object being the odd one out. To this end, we used the softmax function, which has been demonstrated to be suitable for relating representational proximity to similarity in the context of choice models and for estimating generalization behaviour. Finally, the difference between the predicted choice probability and the actual choice served as a model prediction error, which allowed us to adapt the model dimensions in proportion to this error (see Methods for the details on this optimization procedure). The model was trained on 90% of the available trials, and the remaining 10% were later used for an independent assessment of model performance (see below).

A stable and predictive model of behaviourally measured similarity. As expected, due to the sparsity constraint, many of the initial dimensions revealed values close to 0 and were discarded, leaving us with 49 dimensions. We then sorted the dimensions on the basis of the sum of all dimension values across all objects, in descending order. Due to the stochastic nature of the modelling procedure, fitting the model repeatedly may lead to a different embedding and a slightly different number of dimensions. To estimate the stability of the model, we re-ran it 20 times with different random initializations (Methods). Across those models, most dimensions exhibited high reproducibility (Pearson’s $r > 0.9$ in 34/49 dimensions, Pearson’s $r > 0.6$ in 46/49 dimensions), demonstrating that the procedure generated a highly stable and reproducible embedding (see Extended Data Fig. 1 for a plot of the reproducibility of all dimensions). There was a strong correlation between the ranks of the dimensions and the dimension reproducibility (Spearman’s $\rho = 0.75$; $P < 0.001$; randomization test; 95% confidence interval (CI), 0.61–0.85), indicating that the reproducibility of individual dimensions was driven mostly by their overall importance in the model.

Having demonstrated the reproducibility of the model dimensions, we next tested the predictive performance of the model. We first estimated how well we could predict individual choices in the odd-one-out task using trials from the independent test set. To gain an understanding of the best possible prediction any model could achieve for these 1,854 objects given the variation present in the data (the noise ceiling), we additionally sampled 1,000 randomly chosen triplets 25 times and estimated the consistency of choices for each triplet across participants. Averaged across those triplets, the upper limit in fitting individual-trial behaviour from the data was 67.22% (±1.04%). Overall, the model correctly predicted 64.60% (±0.23%) of individual trials in the independent test data (Fig. 2a). This means that the model achieved 92.25% (±1.50%) of the best possible accuracy at predicting behaviour, demonstrating excellent predictive performance at the individual-trial level given the noise in the data.
To evaluate how well the model could predict behaviourally measured similarity, we next generated a fully sampled similarity matrix of 48 diverse objects and compared it with the similarity matrix predicted by our model. Since we had sampled only a fraction of the 1,854 × 1,854 similarity matrix, the test data were insufficient for addressing how well the model could predict behaviourally measured similarity. We therefore used online crowdsourcing to collect two or three behavioural responses for each possible triplet of those 48 objects (43,200 choices) and calculated the choice probabilities for each pair of objects as a measure of their similarity. To estimate the noise in the fully sampled matrix, we calculated the reliability by splitting the behavioural data in half and generating two split-half similarity matrices. We then computed a predicted similarity matrix using our computational model and compared it with both the full similarity matrix and each split. The predicted and measured similarity matrices are depicted in Fig. 2b. Both matrices were highly correlated (Pearson’s $r = 0.90$; $P < 0.001$; randomization test; 95% CI, 0.88–0.91), with the fit of each half again approaching the noise ceiling (first half, $r = 0.87$; second half, $r = 0.88$; reliability, $r = 0.91$), demonstrating that the model was able to accurately reproduce behaviourally measured similarity even with very sparsely sampled data. This result highlights that despite the large number of objects and the complexity of natural stimuli, most of the large-scale representational structure of objects measured through human similarity judgements can be captured by a fairly low-dimensional embedding.

Are the model’s dimensions interpretable? The results so far establish that the model dimensions are reproducible, can be used to accurately generate similarities between pairs of objects, and predict individual behaviour close to the noise ceiling. However, they leave open the degree to which individual model dimensions can be interpreted meaningfully. If the dimensions are interpretable, then the objects with the highest weight in a given dimension should share certain properties that are easy to identify. In Fig. 3, we illustrate the interpretability for a subset of dimensions by displaying the object images with the highest weights along those dimensions. The word clouds illustrate the labels provided by 20 participants to the visual exposure of those dimensions, weighted by their naming frequency. While the responses tended to focus more on extreme examples, they generally exhibited a close correspondence to the dimension labels we generated, which are shown above each set of images (see Extended Data Fig. 2 for the labels and word clouds of all dimensions). For this figure, all images were replaced by images with similar appearance from the public domain. Images used under a CC0 license, from Pixabay: blende12, Beeki, Bernell, PublicDomainPicture, Alexas_Fotos, Katya36, creazionpublicidad, VanVangelis, skeez, DEZALB, Arbohemia, alien0417, Pezibear, BRRT, Neutrini, moritz320, arembowski, Hans, terimakashi0, LAWJR; Pexels: Clovis Cheminot, Burak K; Wikimedia: Paul de Lamerie; Flickr: monstersforsale, U.S. Navy, Fried Dough; PublicDomainPictures: Alex Borland; DVIDSHub: Department of Defense; Peakpx; Pikrepo; PxHere.
such as being metallic or hard, valuable, disgusting, heat related or water related. Finally, some dimensions seem to reflect perceptual properties, such as the roundness of objects, their elongation, flatness, colour, shininess or patterned texture. For later use throughout this manuscript, we assigned intuitive labels to each dimension (such as ‘animal-related/organic’ and ‘colourful’).

To explicitly test this interpretability in naïve observers, we asked 20 laboratory participants (15 female, 5 male) to provide labels for those dimensions, on the basis of viewing objects sorted by their numeric values along each dimension. Since interpretability need not be limited to a single label, we visualized the naming results using word clouds, in which more frequently provided labels are displayed with a larger font. While participants’ descriptions varied and tended to focus more on extreme examples of a dimension, they exhibited a remarkably close correspondence to the labels we had assigned to the dimensions (see Extended Data Fig. 2 for the naming of all 49 dimensions).

Having established the interpretability of object dimensions, we can explore what dimensions a given object is composed of. For that purpose, in Fig. 4 we visualize a range of different objects using circular bar plots (rose plots), where the angle and colour of a petal reflect the object dimension and the length of the petal reflects the degree to which the dimension is expressed in that object. For example, the image of noodles is characterized mostly by being food related, repetitive and stringy. In contrast, the image of a rocket is characterized mostly by being transportation related, flying related, fire related, artificial and shiny. This visualization demonstrates that some dimensions indeed reflect perceptual properties, since they are specific for the chosen object images. They may not show up for a different image of the same object and might have been missed completely if words had been chosen instead of images. In addition, the visualization demonstrates that objects are indeed characterized by a rather small number of dimensions (see below for a quantification).

**Natural object categories as an emergent property of similarity embedding.** To characterize the relative similarity of objects to each other and explore the distribution of dimensions across objects, we combined two common visualization tools. First, we projected the 49-dimensional similarity embedding to 2 dimensions using t-distributed stochastic neighbourhood embedding (dual perplexity, 5 and 30), initialized using metric multidimensional scaling. This approach has been shown to preserve the global similarity structure while providing a higher degree of interpretability at the local similarity level than multidimensional scaling alone. Second, in this two-dimensional plot we visualized each object using rose plots (as in Fig. 4).

The resulting visualization (Fig. 5) reveals several interesting features of the similarity embedding. The global similarity structure seems to highlight the well-known distinctions of animate vs inanimate and natural vs man-made, but it also reveals three differences. First, representations of humans and human body parts are largely separate from animals and closer to the man-made objects, in line with neuropsychological findings and demonstrating an important limitation of simply applying taxonomic relationships for studying mental representations. Second, processed food was found to be more closely related to living and natural objects, acting as an exception to the universality of naturalness as a critical dimension of object representations, again in line with patient data. Third, the weights of dimensions do not reflect binary membership to the categories of, for example, natural and man-made objects. For example, focusing on the dimension ‘artificial/hard’ (dark-blue, rightward-oriented bars), this dimension was most strongly expressed on the right of the graph but became weaker when moving to the left, towards animals, food and natural objects.

In addition to this global structure, many objects formed clusters related to high-level categories (such as animals, tools, vehicles and musical instruments). This indicates that categorization behaviour for many categories may be accounted for by the similarity of objects, a property that has been discussed previously but had not been tested on a large set of natural objects. To test how well natural categories could be predicted by the similarity embedding, we used 18 unique high-level categories identified in the THINGS database and used a cross-validated nearest-centroid classifier to predict category membership for each of the 1,112 objects of those categories. The classifier performed at 86.42% (chance performance, 5.56%), on par with a recent semantic embedding of object meaning (85.97%) that had been trained on billions of words, demonstrating that the dimensions we identified allow the prediction of categorization behaviour for a large number of natural categories.

Finally, the visualization reveals that certain combinations of dimensions are critical for forming different types of categories. Indeed, many subcategories can be explained by defining features: objects with large weights in the animal-related and water-related dimensions are probably sea creatures, while objects with large weights in the plant-related and food-related dimensions are probably vegetables. For example, an abacus can be explained as a combination of several dimensions, such as artificial/hard, wood-related, valuable/special and coarse pattern.

**How many dimensions for an object?** The visualizations in Figs. 4 and 5 suggest that some objects are easy to characterize with a relatively small number of dimensions. This indicates that, while all 49 dimensions are useful for some objects, individual odd-one-out judgements of objects may be predicted accurately with a smaller number of dimensions. Since the model performs close to the noise ceiling at predicting behaviour, we should be able to produce a lower-bound estimate for the number of relevant object dimensions for characterizing individual behaviour and the global similarity structure. To this end, we carried out a dimension-elimination approach. The reasoning behind this approach is that if a dimension does not matter for behaviour, then setting it to 0 should not affect the predictive performance of the model. Therefore, for each object, we set the dimension with the lowest weight to 0, predicted behaviour in the test set, recomputed the similarity matrix and compared how this affected the predictive performance of the model. We then repeated this elimination process until only one dimension was left. Importantly, this procedure does not eliminate entire dimensions from all objects but eliminates different dimensions for each object. If behaviour is driven by a larger number of dimensions than retained, this would be reflected in reduced model performance. The results of these analyses demonstrate that to achieve 95–99% of the performance of the full model in explaining behavioural judgements in the odd-one-out task, a total of 6 to 11 dimensions are required (Fig. 6a), and to explain 95–99% of the variance in the similarity matrix, a total of 9 to 15 dimensions are required (Fig. 6b). Thus, while the representational space of objects can be captured by a comparatively low-dimensional embedding, for judging the similarity of objects, on average humans indeed seem to integrate across a larger number of those dimensions.

---

**Fig. 4 | Illustration of example objects with their respective dimensions, using rose plots.** The length of each petal reflects the degree to which an object dimension is expressed for the purpose of a given object. For display purposes, dimensions with small weights are not labelled. For this figure, all images were replaced by images with similar appearance from the public domain. Images used under a CC0 license, from Pixabay: clangs, Svyatoslav, Nuschk466, Tabajachi, galdosak, GAIMARD; flickr: jimmy_c_1991; peakpx; pikrepo.
Typicality as an emergent property of similarity embedding.

While objects are characterized by several different dimensions, it is unclear to what degree these dimensions merely reflect binary properties of the objects (for example, ‘is an animal’, ‘is a tool’ or ‘is a vehicle’) or rather the degree to which they are present in an object (for example, animacy, manipulability or utility for...
transportation). While we were able to predict the high-level categories of objects from the embedding, we did not test whether the continuous nature of the dimensions was informative about the degree to which a dimension is expressed in an object. The continuous nature of dimensions may be reflected in the typicality of objects within their corresponding categories. Interestingly, this would demonstrate that not only object categories but also typicality can be described as an emergent property of object dimensions. To test whether the numeric value of a dimension reflected typicality, we used online crowdsourcing to collect typicality ratings for words of the 27 high-level categories in the THINGS database. Of those categories, 17 could be related to dimensions of our embedding according to their dimension labels; we therefore tested their correspondence with typicality. The results of these analyses are shown in Fig. 7 and Extended Data Fig. 3. Despite the typicality ratings being based on words and the dimensions on images, 14 out of 17 dimensions revealed a significant positive relationship with typicality scores (Spearman’s $\rho$, $0.26–0.62$; all $P < 0.05$; one-sided; false discovery rate (FDR) corrected for multiple comparisons). These results demonstrate that typicality may be an emergent property of the object dimensions. However, the results also reveal that some dimensions with a weaker relationship do not seem to reflect a purely category-related semantic code but may incorporate other, perhaps perceptual aspects.

Human ratings along model dimensions allow the generation of similarity scores for arbitrary object images. A generative model of object similarity would open the possibility to directly operate on the dimensions rather than having to collect similarity judgements. To what degree can the representational embedding identified from the odd-one-out judgements act as a generative model of object similarity? A simple way to test this idea is to ask participants to provide direct ratings of objects along the dimensions of the embedding31. If it is possible to generate similarity from human ratings of dimensions, this would also serve as a stricter test of their interpretability.

For a set of 20 object images selected at random from the 1,854 objects, we asked 20 laboratory participants (15 female, 5 male) to rate those objects along the 49 object dimensions of the representational embedding (Fig. 8a). Rather than providing them with semantic labels for the dimensions, we showed the participants example images along a continuous rating scale and asked them where along that scale they would place those objects. In the next step, the responses for each dimension were averaged across participants and used in place of the model dimensions to generate a human-predicted similarity matrix. Importantly, since the focus here was to provide proof of concept for the usefulness of the dimensions, the participants had not been trained on this task.

The comparison of similarity from direct dimension ratings and the reference similarity matrix from the model embedding

---

**Fig. 5 | Two-dimensional visualization of the similarity embedding, combining dimensionality reduction (multidimensional-scaling-initialized $t$-distributed stochastic neighbourhood embedding; dual perplexity, 5 and 30; 1,000 iterations) with rose plots for each object (see Fig. 4).** At the global structure level, the results confirm the well-known distinctions of animate vs inanimate and man-made vs natural objects, with some exceptions (see the main text). In addition, the different clusters seem to reflect broader object categories that emerge naturally from object similarity judgements. However, the dimensions are not restricted to those clusters but are expressed to different degrees throughout this representational space. For this figure, all images were replaced by images with similar appearance from the public domain. Images used under a CCO license, from Pixabay: fotoblend, D_Van_Rensburg, kropekk_pl, savagexy, cristinamacia, skeeze, HuaHinTown_Snap, murray0710.
articles.nature.com

Identifying the structure of our internal mental representations is a central goal in the cognitive sciences. For the domain of natural objects, this may seem particularly challenging given the high complexity of our visual world, which contains thousands of objects with a seemingly countless number of possible object properties. Here, using a triplet odd-one-out task on a wide range of object images, we demonstrated that it is possible to characterize the similarity structure and individual human behavioural judgements with a low-dimensional representational embedding learned directly from human choice behaviour. The model revealed 49 meaningful object dimensions, each being interpretable with respect to the perceptual and conceptual properties of those objects, reflecting basic perceptual properties of shape, colour and texture as well as higher-level properties such as taxonomic membership, function and value. The embedding allowed the prediction of other forms of behaviour, including high-level categorization and typicality judgements. By demonstrating that participants can use these dimensions to generate object similarity scores, these results open the avenue towards a generative model of object similarity judgements. Importantly, the resulting large-scale similarity matrix based on our representational embedding can act as a basis for testing formal computational models of categorization and category learning in the domain of natural objects.

Discussion

Being able to characterize mental representations of objects with a low-dimensional embedding is surprising, given the objects’ high degree of perceptual variability and our broad semantic knowledge of them. Indeed, popular semantic feature production norms have revealed thousands of binary features that participants name when asked about their explicit knowledge of objects. Rather than attempting to capture all the details of our semantic knowledge of objects with binary properties, our results demonstrate that it is possible to achieve high predictive performance using only a small number of interpretable, continuous dimensions. It may be possible to generate these binary properties from the continuous dimensions in our model, which would demonstrate that the implicit judgements in the odd-one-out task capture much of the explicit semantic knowledge of objects, but a general test of this idea would require the creation of feature production norms for the 1,854 objects used in the creation of the embedding. However, even if such norms were created, there are two reasons that their predictions may be limited. First, the dimensions revealed in this work are focused around the properties most relevant for discriminating among different objects, while feature production norms would probably contain much more (often idiosyncratic) information than required for those distinctions. Second, when using object words and an explicit feature-naming task, participants often omit critical features. By using image objects and an implicit, non-verbal task, it is possible to capture perceptual dimensions of objects with a representational embedding that might otherwise be missed.

In contrast to traditional data-driven approaches that identify multidimensional feature spaces using dimensionality reduction techniques such as multidimensional scaling, factor analysis and additive clustering, for the present model we made two assumptions that support the interpretability of dimensions, motivated by the observation of how objects are typically characterized: (1) dimensions are sparse (that is, each object carries only some dimensions but not others) and (2) dimensions are positive (that is, each object is characterized by a combination of dimensions that are present to a certain degree and that add up without cancelling each other out). By incorporating these assumptions, our model not only yields interpretable dimensions but also reflects a blend between two common model families used to characterize objects: dimensional models that assume continuous dimensions, and featural models that assume the presence and absence of (mostly binary) object properties. Analyses of category-related typicality judgements demonstrate that the continuous nature of the dimensions is informative as to the degree to which these dimensions are expressed in objects, demonstrating that continuous dimensions allow us to generalize beyond binary categorical assignment of semantic attributes (for example, ‘is animate’). In addition, whereas the traditional pairwise assessment of similarity...
typically neglects the importance of object context\textsuperscript{41}, by using a triplet odd-one-out task, this type of embedding in principle allows the generation of object similarity for arbitrary contexts imposed by focusing on a chosen subset of objects (such as animals). The degree to which the embedding carries such fine-grained information will need to be tested in future studies.

Mathematically, there are an infinite number of possible ways in which object representations can be characterized by a set of dimensions. Identifying a broad range of meaningful and predictive dimensions with a bottom-up, data-driven model offers a systematic approach for the identification of meaningful dimensions, complementing traditional top-down, theory-driven approaches. Ultimately, however, further studies are required to validate the specificity of different dimensions in this model and link them to representations in the human brain. One intriguing prediction of our model is that specific deficits in recognizing objects found in patients with focal lesions may be tied more to specific dimensions of training participants to generate dimension ratings could rely on those ~1% of the objects would strongly bias the results, and collecting ratings from a different set of objects would have required prohibitive in the context of this study. Future studies with the goal of training participants to generate dimension ratings could rely on a separate set of objects for relating predicted similarity with measured similarity.

Finally, the prediction of similarity from direct ratings of dimensions was based on 20 objects that were part of creating the original embedding, which may slightly overestimate the ability to generate similarity from dimension ratings. However, given the large number of objects used in this study, we believe it to be unlikely that those ~1% of the objects would strongly bias the results, and collecting ratings from a different set of objects would have required generating, characterizing and testing a different model, which was prohibitive in the context of this study. Future studies with the goal of training participants to generate dimension ratings could rely on a separate set of objects for relating predicted similarity with measured similarity.

The approach proposed in this study opens the avenue for many related questions. To what degree are the dimensions shared between different individuals\textsuperscript{42,43} and how are they affected by gender, age, culture, education, other socio-demographic factors and individual familiarity with the objects? To what extent do the representations depend on the exact task, and can other similarity tasks evoke similarly fine-grained representations?\textsuperscript{44,45} What are the representational dimensions in other domains, such as words, and different domains? Addressing these questions will be important for a comprehensive understanding of mental representations of objects across people and different domains.
Methods

Participants. A total of 5,983 workers from the online crowdsourcing platform Amazon Mechanical Turk participated in the triplet odd-one-out experiments, which consisted of the creation of the fully sampled matrix of 48 objects (121 workers, 100 after exclusion; 3,159 female, 2,092 male, 19 other, 31 not reported), and the 1,000 randomly chosen triplets used for the estimation of a noise ceiling (336 workers, 325 after exclusion; 156 female, 103 male, 66 not reported). In addition, a total of 337 workers (no exclusions; 198 female, 131 male, 8 not reported) participated in the creation of the typicality norms. All workers were located in the United States, and worker ages were not assessed. For the odd-one-out task, workers were excluded if they exhibited overly fast responses in at least 5 sets of 20 trials (the speed cut-off was 25% or more responses <800 ms and 50% or more responses <1,100 ms) or if they carried out at least 200 trials and showed overly deterministic responses (>40% of responses in one of the three odd-one-out positions; expected value, 33%). All workers provided informed consent. The number of trials (and consequently the sample size) was determined on the basis of feasibility and available resources. The online research was approved by the Office of Human Research Subject Protection and conducted following all relevant ethical regulations, and the workers were compensated financially for their time.

In addition, 20 laboratory participants (15 female, 5 male; mean age, 26.25; s.d., 6.39; range, 19–41) took part in the dimension labelling and the dimension rating experiment. All laboratory participants provided written informed consent and were compensated financially for their time. No statistical methods were used to predetermine the sample sizes. The laboratory experiments were carried out following all relevant ethical regulations and rules of the National Institutes of Health Institutional Review Board (NCT00001360).

Object images and odd-one-out task procedure. The 1,854 images of objects used in this study were the reference images that had been used previously for validating the concepts of the THINGS database17. The images depict objects embedded in a natural background and were all cropped to square size, except a small number of images that didn’t fit into a square and that were padded with white background on both sides. Importantly, the validation task of the THINGS database demonstrated that the objects in the 1,854 images were generally nameable—that is, it can be assumed that most participants were sufficiently familiar with the objects to be able to name them. The triplet odd-one-out task was carried out in sets of 20 trials, and the workers could choose how many sets they would like to take part in. On each trial, the workers were shown three object images side by side in a browser window and were asked to report the image that was the least similar to the other two. The workers were told that they should focus their judgement on the object, but to minimize bias they were not given additional constraints as to the strategy they should use. In addition, the workers were instructed that if they did not recognize the object, they should base their judgement on their best guess of what the object would fall along all 49 model dimensions. Rather than providing the participants with dimension labels, the rating scale was spanned by example images along the currently rated dimension (in this example, dimension 1, artificial/hard). Results for the 20 tested objects revealed a good reconstruction of object similarity by dimension ratings when comparing it with the similarity predicted from the embedding that served as a reference (Pearson’s $r = 0.85$; $P < 0.001$; randomization test; 95% CI, 0.80–0.89). These results further support the idea that dimensions are interpretable and that they can be used to directly generate object similarities. For this figure, all images were replaced by images with similar appearance from the public domain. Images from PublicDomainPictures: Jean Beaufort; Pixabay: jackmac34, TheDigitalArtist, Svyatoslav, UmerSaud, maim6, kalhh, Bruno, AutoPhotography, ingagestralia, ivakohl, endlesswatts, PixelMission, Brett_Hondow, freestocks-photos, MikeGoad, MabelAmber, GuilleNet; Flickr: craig cannon; Peakpx; Pikrepo.
could be. The workers responded with a mouse click on the respective image, which initiated the next trial after an intertrial interval of 500 ms. Each object triplet and the order of stimuli were chosen randomly, but such that after data collection each cell in the 1,854 × 1,854 similarity matrix had been sampled at least once.

To yield a diverse set of objects for the fully sampled similarity reference dataset, the 48 objects were chosen by carrying out spectral clustering on publicly available 300-dimensional sense vectors of all 1,854 objects\(^6\) with 48 clusters and by choosing one object per cluster randomly.

Details of the computational modelling procedure. The model and preliminary results were presented previously at a conference\(^7\). The model was implemented as a computational graph in TensorFlow (version 1)\(^\dagger\). Each triplet was encoded using three one-hot vectors (length, 1,854), and each vector was linked to 90 latent dimensions, but with weights replicated across all three vectors. The 1,854 × 90 weights were initialized randomly (range, 0–1). Note that initializing the model with 200 dimensions led to a very similar model performance and final number of dimensions (prediction accuracy of test set odd-one-out choices, 64.70%; dimensions, 50). The dot product was chosen as a basis for proximity for computational reasons, but using the Euclidean distance led to similar performance (prediction accuracy of test set odd-one-out choices, 64.69%; dimensions, 57). The objective of the model optimization consisted of the cross-entropy, which was the logarithm of the softmax function, and a regularization term based on the L1 norm to encourage sparsity:

$$\sum \log \left( \frac{\exp(x_{ij})}{\sum \exp(x_{ij})} \right) + \lambda \sum |x_{ij}|$$

where \(x\) corresponds to an object vector; \(i\) and \(j\) to the indices of the current triplet; \(n\) to the number of triplets; and \(m\) to the number of objects. The regularization parameter \(\lambda\), which controls the trade-off between sparsity and model performance, was determined using cross-validation on the training set (\(\lambda = 0.008\)). In addition to sparsity, the optimization was constrained by strictly enforcing weights in the embedding X to be positive. The minimization of this objective was carried out using stochastic gradient descent as implemented in the Adam algorithm\(^8\) using the default parameters and a minibatch size of 100 triplets. After the optimization was complete, dimensions for which the weights of all objects were smaller than 0.1 were removed, leaving us with 49 dimensions. Empirically, the largest maximum weight of all included dimensions was 0.03, while the smallest maximum weight of all included dimensions was 1.38.

The dimensions were sorted in descending order by the sum of their weights across objects.

Computation of the similarity matrix from embedding. We defined object similarity in the triplet odd-one-out task as the probability \(p(i|j)\) of the participants choosing objects \(i\) and \(j\) to belong together, irrespective of context. Therefore, to compute similarity from the learned embedding for all 1,854 objects, we created all predicted choices for all possible \(1 - \binom{1,000}{3}\) triplets and calculated the mean choice probability for each pair of objects. For the fully sampled similarity matrix of 48 objects used for testing the performance of the model at predicting object similarity (Fig. 2), we created a different similarity matrix that was constrained only by this subset of 48 objects.

Reproducibility of embedding dimensions. Due to the stochasticity of the optimization algorithm, each time the model is re-run, we will probably end up with a different set of dimensions. To determine the stability of each dimension in our 49-dimensional embedding, we re-ran the model 20 times, each time with a different random initialization. Next, we correlated each of the 49 original dimensions with all dimensions of one of the 20 reference embeddings and chose the best-fitting dimension across all correlations as the closest match. We then applied a Fisher \(z\)-transform to the correlations, averaged them across all 20 reference embeddings and inverted the Fisher \(z\)-transform to get a mean reliability for each dimension across all 20 embeddings. While the resulting comparison may exhibit a slightly positive bias due to choosing the best fit, a split-half cross-validation between all objects demonstrated nearly indistinguishable results (maximum difference \(r = 0.01\)).

Category prediction. The categorization performance of the representational embedding was tested on 18 of the 27 categories in the THINGS database. Objects that were members of multiple categories were removed. Of the 9 categories that were removed, 7 were subcategories of other categories (for example, ‘vegetable’ in ‘food’) or had fewer than ten objects after the removal of non-unique objects. The remaining 18 categories were as follows: animal, body part, clothing, container, electronic device, food, furniture, home decor, medical equipment, musical instrument, office supply, part of car, plant, sports equipment, tool, toy, vehicle and weapon. These categories comprised 1,112 objects. The classification was carried out using leave-one-object-out cross-validation. For training, centroids for all 18 categories were computed by averaging the 49-dimensional vectors of all objects in each category, excluding the left-out object. The membership of this remaining object was then predicted by the smallest Euclidean distance to each centroid. This procedure was repeated for all 1,112 objects, and the prediction accuracy was averaged. For the corresponding analysis with a semantic embedding, we used publicly available 300-dimensional sense vectors\(^6\).

Typicality ratings. Typicality ratings were collected for all 27 object categories in the THINGS database, with the goal of including them as metadata for the database. However, for the purpose of this study, we focused on the 17 categories for which the labels indicated a relationship of dimensions with specific object categories. Typicality ratings were collected by asking workers to rate the typicality of an object as belonging to a given category, using a Likert scale from 0 to 10. Each rating was collected 40 times. To improve the comparability of the use of the Likert scale, each worker’s responses were z-scored before they were merged with the other responses.

Dimension-naming task and construction of word clouds. In the dimension-naming task, the laboratory participants were asked to provide labels for different dimensions after inspecting them. This was achieved by showing them example object images along a continuous scale for a given dimension, comparable to the procedure described above. The participants could further inspect the dimensions by clicking on example objects, which would reveal more object images in this range. The exact object images shown varied between dimensions. The participants were instructed not to focus exclusively on the top of the scale but to take the entire distribution of objects into account. After having studied a dimension, they were allowed to provide one verbal label for the dimension. The word clouds displaying the participants’ responses according to the frequencies of provided object labels were constructed using the function wordcloud in MATLAB (Mathworks), using the default parameters.

Object dimension rating task. The object dimension rating task was carried out after the dimension-naming task. The participants were shown a reference image on the top and were asked to rate where they would place the object along a Likert scale, according to the meanings of the dimensions explored previously (Fig. 8a). The participants were recommended to take two of the seven shown levels (demarcated by the object images) and judge whether the given object was better characterized by one or the other level. The rating task was carried out for all 49 dimensions sequentially, on images of the following 20 objects chosen randomly from the set of 1,854 objects: bazooka, bib, crowbar, crumb, flamingo, handbrake, hearse, keyhole, palm tree, scallion, sleeping bag, spider web, splinter, staple gun, suitcase, syringe, tennis ball, woman, workbench and wrench. Since more dimensions were at or close to zero, the scale included a short range with all zero values (“not at all”). Furthermore, to improve the discriminatory power of the scale, the dimension values were converted to percentiles, with all percentiles lower than 20% set to 0.2. After the ratings had been collected, the percentiles were converted back to their respective continuous values along the dimensions and averaged across participants. For better comparability to the original dimensions, the dimension values were scaled so that their minimum rating corresponded to zero. The object dimensions were then treated as new embedding dimensions for the 20 objects, and object similarity was calculated for them according to the procedure described above. Finally, the similarity matrix generated from the object ratings was compared with the similarity matrix from the full model.

Statistical analyses and Cls. Unless indicated otherwise in the text, the statistical analyses were conducted using classical parametric statistical tests and, when required, corrected for multiple comparisons using FDR. The data distribution was assumed to be normal, but this was not formally tested. All tests were two-tailed unless denoted otherwise. Non-parametric randomization tests on correlations between the predicted and measured similarity matrices were conducted by creating 100,000 similarity matrices on the basis of randomly shuffling object labels, re-running the correlation with the measured similarity matrix and calculating P values as the percentage of permutations reaching or exceeding the true similarity. The error bars reflect 95% CIs and were created on the basis of the standard error of the mean or, when no distribution was available, the standard deviation of 1,000 bootstrap samples.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability
The learned embedding, triplet odd-one-out behavioural data for testing model performance, typicality scores, participant-generated dimension labels and dimension ratings are available at https://osf.io/z2784. The benchmark data used for training the model are available from the corresponding author upon request.

Code availability
To reproduce the relevant analyses and figures, the relevant MATLAB scripts and functions are available at https://osf.io/z2784. The computational modelling code to create an embedding is available from the corresponding author upon request.
Articles NaTurE HumaN BeHaviOur

31. Nosofsky, R. M., Sanders, C. A., Meagher, B. J. & Douglas, B. J. Toward the development of a feature-space representation for a complex natural category. Behav. Res. Methods 50, 530–556 (2018).

32. Nosofsky, R. M., Sanders, C. A., Meagher, B. J. & Douglas, B. J. Search for the missing dimensions: building a feature-space representation for a natural-science category domain. Comput. Brain Behav. 3, 13–33 (2020).

33. Keil, F. C. Constraints on knowledge and cognitive development. Psychol. Rev. 88, 187–227 (1981).

34. Shepard, R. N. The analysis of proximities: multidimensional scaling with an unknown distance function. Psychometrika 27, 125–140 (1962).

35. Torgerson, W. S. Multidimensional scaling: I. Theory and method. Psychometrika 17, 401–419 (1952).

36. Thurstone, L. M. Multiple factor analysis. Psychol. Rev. 38, 406–427 (1931).

37. Strelau, J., Logan, C. G., Frank, R. J. & Damasio, A. R. Explaining category-related effects in the retrieval of conceptual and lexical knowledge for concrete entities: operationalization and analysis of factors. Neuropsychologia 35, 1329–1339 (1997).

38. Shepard, R. N. & Arabie, P. Additive clustering: representation of similarities as combinations of discrete overlapping properties. Psychol. Rev. 86, 27–123 (1979).

39. Navarro, D. J. & Lee, M. D. Common and distinctive features in stimulus similarity: a modified version of the contrast model. Psychon. Bull. Rev. 11, 961–974 (2004).

40. Carlson, T. A., Ritchie, J. B., Kriegeskorte, N., Durvasula, S. & Ma, J. Reaction times for object categorization is predicted by representational distance. J. Cogn. Neurosci. 26, 132–142 (2013).

41. Yee, E. & Thompson-Schill, S. L. Putting concepts into context. Psychon. Bull. Rev. 23, 1015–1027 (2016).

42. Charest, I., Kievit, R. A., Schmitz, T. W., Deca, D. & Kriegeskorte, N. Unique semantic space in the brains of each beholder predicts perceived similarity. Proc. Nat. Acad. Sci. USA 111, 14565–14570 (2014).

43. De Haas, B., Jakovidas, A. L., Schwarzkopf, D. S. & Gernfurther, K. R. Individual differences in visual salience vary along semantic dimensions. Proc. Nat. Acad. Sci. USA 116, 11687–11692 (2019).

44. Peterson, J. C., Abbott, J. T. & Griffiths, T. L. Evaluating (and improving) the correspondence between deep neural networks and human representations. Cogn. Sci. 42, 2648–2669 (2018).

45. Rajalakshmi, R. et al. Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. J. Neurosci. 38, 7255–7268 (2018).

46. Boaz, K. M., Kriegeskorte, N., Stors, K. R. & Mur, M. Deep convolutional neural networks ouperform feature-based but not categorical models in explaining object similarity judgments. Front. Psychol. 8, 1726 (2017).

47. Jordan, M. C., Galliand, T., Ellis, C. T., Beckage, N. & Cohen, J. D. Context matters: recovering human semantic structure from machine learning analysis of large-scale text corpora. Preprint at arXiv https://arxiv.org/abs/1910.06954 (2019).

48. Bauer, A. J. & Just, M. A. in The Oxford Handbook of Neurolinguistics (eds De Zubicaray, G. I. & Schiller, N. O.) 518–547 (Oxford Univ. Press, 2019).

49. Binder, J. R. et al. Toward a brain-based componential semantic representation. Cogn. Neuropsychol. 33, 130–174 (2016).

50. Huth, A. G., Nishimoto, S., Vu, A. T. & Gallant, J. L. A continuous semantic space describes the representation of thousands of object and action categories across the human brain. Neuron 76, 1210–1224 (2012).

51. Cicchini, R. M., Kriegeskorte, N., Joosik, K. M., van den Bosch, J. I. & Charest, I. The spatiotemporal neural dynamics underlying perceived similarity for real-world objects. NeuroImage 194, 12–24 (2019).

52. Bankson, B. B., Hebart, M. N., Groen, I. I. A. & Baker, C. I. The spatiotemporal dynamics of visual object representation across the human visual cortex. Neuron 108, 1680–1690 (2020).

53. Acknowledgements

We thank A. Corriveau for help with the data collection in the laboratory experiment. S. Nosofsky and J. Perkstein for help in editing the paper. We also thank the Max Planck Society for funding public-domain images for this publication, and I. Charest, B. Love, A. Martin and P. McClure for helpful discussions and/or comments on the manuscript. This research was supported by the Intramural Research Program of the National Institutes of Health (grant nos ZIA-MH-002909 and ZIC-MH00298), under National Institute of Mental Health Clinical Study Protocol 93-M-1070 (NCT0001360), and by a research group grant awarded to M.N.H. by the Max Planck Society. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.
Author contributions
M.N.H. and C.I.B. conceived and designed the study. M.N.H. collected the data. C.Y.Z., M.N.H. and F.P. designed the computational model. M.N.H., C.Y.Z. and F.P. analysed the data. M.N.H., C.I.B., F.P. and C.Y.Z. wrote the manuscript and provided critical revisions.

Competing interests
The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41562-020-00951-3.

Supplementary information is available for this paper at https://doi.org/10.1038/s41562-020-00951-3.

Correspondence and requests for materials should be addressed to M.N.H.

Peer review information
Primary Handling Editor: Marike Schiffer.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This is a U.S. government work and not under copyright protection in the U.S.; foreign copyright protection may apply 2020.
Extended Data Fig. 1 | Reproducibility of dimensions. Reproducibility of dimensions in the chosen 49-dimensional embedding across 20 random initializations (see Extended Data Fig. 2 for a list of all dimension labels). Shaded areas reflect 95% confidence intervals.
Extended Data Fig. 2 | Labels and word clouds for all 49 model dimensions. Labels for all 49 dimensions, with respective word clouds reflecting the naming frequency across 20 participants. The dimensions appear to reflect both perceptual and conceptual properties of objects. A visual comparison between labels and word clouds indicates a generally good agreement between participant naming and the labels we provided for the dimensions.
| Dimension name                   | Category name | Number of objects in category | Spearman’s $\rho$ (90% CI) | $p$-value (uncorrected) | $p$-value (FDR -corrected) |
|---------------------------------|---------------|-------------------------------|----------------------------|------------------------|--------------------------|
| weapon / danger-related         | weapon        | 48                            | 0.62 (0.43 -0.76)          | < 0.001                | < 0.001                  |
| transportation / dynamic        | vehicle       | 70                            | 0.62 (0.45 -0.75)          | < 0.001                | < 0.001                  |
| furniture-related               | furniture     | 38                            | 0.61 (0.45 -0.74)          | < 0.001                | < 0.001                  |
| electronic / technology         | electronic device | 74                     | 0.60 (0.45 -0.71)          | < 0.001                | < 0.001                  |
| animal-related                  | animal        | 177                           | 0.58 (0.48 -0.67)          | < 0.001                | < 0.001                  |
| sport-related                   | sports equipment | 63                         | 0.53 (0.34 -0.68)          | < 0.001                | < 0.001                  |
| clothing-related                | clothing      | 108                           | 0.52 (0.39 -0.62)          | < 0.001                | < 0.001                  |
| fluid-related / drink-related   | drink         | 19                            | 0.46 (0.04 -0.74)          | 0.026                  | 0.034                    |
| food-related                    | food          | 294                           | 0.42 (0.33 -0.50)          | < 0.001                | < 0.001                  |
| child/toy-related               | toy           | 34                            | 0.37 (0.04 -0.63)          | 0.016                  | 0.024                    |
| instrument-related              | musical instrument | 33                         | 0.35 (0.08-0.58)          | 0.023                  | 0.033                    |
| body part-related               | body part     | 34                            | 0.33 (0.04 -0.56)          | 0.030                  | 0.036                    |
| medicine-related                | medical equipment | 27                     | 0.32 (-0.09-0.64)         | 0.052                  | 0.059                    |
| tool-related                    | tool          | 107                           | 0.28 (0.14 -0.41)          | 0.002                  | 0.004                    |
| container-related / hollow      | container     | 105                           | 0.26 (0.10 -0.40)          | 0.004                  | 0.007                    |
| insects / disgusting            | insect        | 17                            | 0.18 (-0.25 -0.55)         | 0.245                  | 0.261                    |
| plant-related / green           | plant         | 47                            | -0.07 (-0.32 -0.19)        | 0.688                  | 0.688                    |

**Extended Data Fig. 3 | Category-typicality correlations.** Detailed results of inferential statistical analyses correlating category-related dimensions with typicality of their category. One-sided $p$-values were generated using randomization tests and were controlled for false discovery rate (FDR) across multiple tests. 90% confidence intervals were used to complement one-sided tests.
Extended Data Fig. 4 | Model performance and dimensionality as a function of training data size. Model performance and dimensionality varied as a function of the amount of data used for training the model. Models were trained in steps of 100,000 trials. Six models with random initialization and random subsets of data were trained per step and all models applied to the same test data as in the main text, making it a total of 78 trained models. For each step, computation of up to two models did not complete on the compute server for technical reasons, making the total between 4 and 6 models per step. Results for each individual model and the average for each step are shown in the Figure. a. Model performance was already high for 100,000 trials as training data but increased with more data, saturating around the final model performance. b. Dimensionality increased steadily with the amount of training data.
### Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](https://www.nature.com/authors) and the [Editorial Policy Checklist](https://www.nature.com/authors/editorial_policies).

### Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

- **n/a** Confirmed
- [ ] The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- [ ] A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- [ ] The statistical test(s) used AND whether they are one- or two-sided
  - Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- [ ] A description of all covariates tested
- [ ] A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- [ ] A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- [ ] For null hypothesis testing, the test statistic (e.g. $F$, $t$, $r$) with confidence intervals, effect sizes, degrees of freedom and $P$ value noted
  - Give $P$ values as exact values whenever suitable.
- [ ] For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- [ ] For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- [ ] Estimates of effect sizes (e.g. Cohen’s $d$, Pearson’s $r$), indicating how they were calculated

Our web collection on [statistics for biologists](https://www.nature.com/articles/ncb2834) contains articles on many of the points above.

### Software and code

Policy information about [availability of computer code](https://www.nature.com/authors/withdata).

**Data collection**

- Online experiments: custom HTML/CSS/Javascript code
- Laboratory experiments: Matlab code (version R2016b) / Psychophysics toolbox V3

**Data analysis**

- Computational model: TensorFlow Version 1
- Other data analyses: Matlab (versions R2016b and R2019b)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](https://www.nature.com/authors/withdata) for further information.

### Data

Policy information about [availability of data](https://www.nature.com/authors/withdata).

All manuscripts must include a [data availability statement](https://www.nature.com/articles/sdata201615). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The learned embedding, triplet odd-one-out behavioral data for testing model performance, typicality scores, participant-generated dimension labels, and dimension ratings will be made available on OSF upon acceptance. The behavioral data used for training the model are available upon request and will be submitted separately to a large-scale data publication journal for general public availability.
Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

- Life sciences
- Behavioural & social sciences
- Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

| Study description | Quantitative experimental research |
| Research sample   |  
- Online experiments: 6,063 workers on crowdsourcing platform Amazon Mechanical Turk with location USA
- Age information not available for online experiments, gender: 3,559 female, 2,380 male, 19 other, 105 not reported
- Laboratory experiments: 20 volunteers from Laboratory of Brain & Cognition database (15 female, 5 male, mean age: 26.25)
- Samples are not representative of the general population |
| Sampling strategy | Sampling strategy:
- Online experiments: Convenience sampling; workers chose whether to participate in the study but were assigned randomly to different trials
- Laboratory experiments: Convenience sampling; participants were sampled based on availability
Sample size determination:
- Large-scale triplet experiment: determined based on available financial resources
- Triplet validation experiment (48 objects): determined by number of available trials / time per worker
- Noise ceiling experiment: determined to achieve 25 repeats / 1,000 triplets
- Typicality experiment: determined to achieve 40 ratings / object
Justification for sample sizes:
- Online experiments: Yield excellent predictive performance in independent test data, most comparisons not carried out within-participant but pooled across participants (fixed effects), for typicality experiment comparable or larger than existing ratings
- Laboratory experiments: based on previous studies, for relevant quantitative comparisons also pooled across participants |
| Data collection | Online experiments: Data collection in web browser using online platform, no interaction between researcher and participant
Laboratory experiments: Data collection on laptop with large external monitor, researcher was blind as to the goals of the experiments (validation of dimension labels / prediction of similarity) |
| Timing | Triplet validation experiment (48 objects): Start date: Nov 30 2017, Stop date: Dec 02 2017
Large-scale triplet experiment: Start date: Feb 05 2018, Stop date: Jul 12 2018
Noise ceiling experiment: Start date: Sep 17 2018, Stop date: Sep 17, 2018
Laboratory experiments: Start date: Aug 24 2018, Stop date: Oct 19, 2018
Typicality experiment: Start date: Jun 07 2019, Stop date: Jun 09 2019 |
| Data exclusions | Triplet validation experiment (48 objects): 21 excluded
Large-scale triplet experiment: 225 excluded
Noise ceiling experiment: 11 excluded
Typicality experiment: no exclusions
Laboratory participants: no exclusions |
| Non-participation | n.a. |
| Randomization | n.a. |

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.
Materials & experimental systems

- Antibodies
- Eukaryotic cell lines
- Palaeontology
- Animals and other organisms
- Human research participants
- Clinical data

Methods

- ChIP-seq
- Flow cytometry
- MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics: see above

Recruitment:
Online experiments: Workers from Amazon Mechanical Turk (US-location, 95% acceptance rate) could choose to participate.
Laboratory experiments: Participants were contacted as members of the Laboratory of Brain and Cognition participant database

Ethics oversight:
Online experiments: Office of Human Research Subject Protection (OHSRP)
Laboratory experiments: National Institutes of Health (NIH) Institutional Review Board (NCT00001360)

Note that full information on the approval of the study protocol must also be provided in the manuscript.