Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Dimensions underlying human understanding of the reachable world

Emilie L. Josephs [a,c,*], Martin N. Hebart [b], Talia Konkle [c]

[a] Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, USA
[b] Vision and Computational Cognition Group, Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany
[c] Psychology Department, Harvard University, Cambridge, USA

ARTICLE INFO

Keywords:
Reachspaces
Scenes
Representational space
Similarity judgments
Behavior

ABSTRACT

Near-scale environments, like work desks, restaurant place settings or lab benches, are the interface of our hand-based interactions with the world. How are our conceptual representations of these environments organized? What properties distinguish among reachspaces, and why? We obtained 1.25 million similarity judgments on 990 reachspace images, and generated a 30-dimensional embedding which accurately predicts these judgments. Examination of the embedding dimensions revealed key properties underlying these judgments, such as reachspace layout, affordance, and visual appearance. Clustering performed over the embedding revealed four distinct interpretable classes of reachspaces, distinguishing among spaces related to food, electronics, analog activities, and storage or display. Finally, we found that reachspace similarity ratings were better predicted by the function of the spaces than their locations, suggesting that reachspaces are largely conceptualized in terms of the actions they support. Altogether, these results reveal the behaviorally-relevant principles that structure our internal representations of reach-relevant environments.

1. Introduction

While we may never know how a raven is like a writing desk (Carroll, 2000)'[1] we can confidently articulate how a writing desk is like a library desk, and not like a spaceship control panel. What knowledge supports this judgment? Judgments of similarity emerge in part because the world is structured and predictable: entities can be divided into types, and entities of the same type will share a set of properties.(Mervis & Rosch, 1981; Shepard & Arabie, 1979) To date, extensive research has uncovered much of the organization of object and scene properties. (Caramazza & Shelton, 1998; Greene, Baldassano, Esteva, Beck, & Fei-Fei, 2016; Greene & Oliva, 2006; Greene & Oliva, 2009; Hebart, Zheng, Pereira, & Baker, 2020; Huth, Nishimoto, Vu, & Gallant, 2012; Konkle & Oliva, 2012; McRae, Cree, Seidenberg, & McNorgan, 2005; Murphy, 2004; Oliva & Torralba, 2006; Patterson, Xu, Su, & Hays, 2014; Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976) However, the representations underlying the rich, near-scale environments in which we perform hand-based actions have only recently begun to be explored. (Josephs & Konkle, 2019; Josephs & Konkle, 2020; Previc, 1998)

Consider the desktop environment where you type an email, the tabletop where you enjoy a meal, or the kiosk where you check in for a flight. These reach-relevant spaces (hereafter “reachspaces”) are highly behaviorally-relevant environments, which support many of our hand-based tasks and activities and form the backdrop to many of our day-to-day behaviors (see Fig. 1 for examples). They differ from both singleton objects and navigable-scale scenes: they encompass spatial extent and multiple objects, but they require coordination of the hands among graspable objects, rather than transportation of the body through an enclosing space. Recently, evidence has emerged that reachspace images have distinct visual statistics from object or scene images (Josephs & Konkle, 2019; Torralba & Oliva, 2002) and elicit distinct topographies of activity in the brain, with particularly strong recruitment of parietal regions.(Bartolo et al., 2014; Josephs & Konkle, 2019) However, the factors that structure human knowledge of these reach-relevant environments have not been systematically mapped.

Mapping the conceptual structure of a stimulus domain makes explicit what properties of the domain are relevant to human behavior. This mapping can reveal the *dimensions, or* stimulus attributes, that most distinguish among entities in the domain. Such dimensions have been proposed to form the mental axes along which foundational cognitive

* Corresponding author at: Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 32 Vassar St, Cambridge 02139, USA.
E-mail addresses: ejosephs@mit.edu (E.L. Josephs), hebart@cbs.mpg.de (M.N. Hebart), tkonkle@fas.harvard.edu (T. Konkle).

[1] In Alice's Adventures in Wonderland by Lewis Carroll(Carroll, 2000), the Mad Hatter poses a riddle: “Why is a raven like a writing desk?”. Famously, he never provides the answer, and generations of readers have been left to guess for themselves.

https://doi.org/10.1016/j.cognition.2023.105368
Received 28 March 2022; Received in revised form 20 December 2022; Accepted 9 January 2023
Available online 13 January 2023

operations such as categorization and generalization operate.(Goldstone, 1994; Landau, Smith, & Jones, 1988; Mervis & Rosch, 1981; Rosch et al., 1976; Tversky, 1989) It can also reveal the *major classes* that humans implicitly identify within the domain. These major joints in representations provide hypotheses for the aspects of a domain that organize cognitive and neural representations. To date, insights about reachable environments come from studies in applied areas, such as ergonomics, human factors engineering, and environmental psychology. These have suggested some properties which distinguish among reachspaces but have generally explored this question within a very narrow scope. For example, some distinctions have been proposed based on the action demands of the space, including digital versus traditional media workspaces,(Morris, Brush, & Meyers, 2007) or workspaces which support precision work versus strength-based work.(Das & Sengupta, 1996) Other distinctions have been proposed based on the people using the space (e.g. experts versus novices (Kirsh, 1995); individual versus collaborative work(Potvin, Swindells, Tory and Storey, 2012; Scott, Carpendale and Inkpen, 2004). However, these divisions have been explored piecemeal, by making a-priori distinctions within a prescribed kind of workspace, for example by testing for differences between personal and collaborative spaces within the circumscribed category of digital office workstations. Thus, it remains an empirical question how our knowledge of the broader reachable world is structured, what dimensions this structure is based on and whether distinct classes can be identified in a data-driven manner from a large and comprehensive sample of everyday environments.

One way to understand the structure of internal representations is to probe the similarity among many exemplars of a concept.(Edelman, 1998; Shepard, 1987) In representational similarity analysis, the similarity (or dissimilarity) among items is conceptualized geometrically as the distance between them in a multi-dimensional feature space and is often expressed as a matrix of pairwise distances.(Dobs, Isik, Pantazis, & Kanwisher, 2019; Groen et al., 2018; Jozwik, Kriegeskorte, Storrs, & Mur, 2017; King, Groen, Steel, Kravitz, & Baker, 2019; Kriegeskorte, Mur, & Bandettini, 2008) These similarity measurements can be leveraged to discover a low-dimensional embedding space for a set of concepts, revealing the dimensions along which concepts vary. Here, we use large-scale crowdsourcing and computational modeling to reveal the similarity structure underlying our knowledge of reachable environments and derive key properties underlying this structure.

In the present work, we collected 1.25 million behavioral similarity judgments on a set of 990 images of reach-relevant environments and used computational modeling to capture the representational structure of these judgments. Broadly, we find that a 30-dimensional space can account for the similarity structure in the judgments and that this space can be divided into four distinct classes. Additionally, we find that

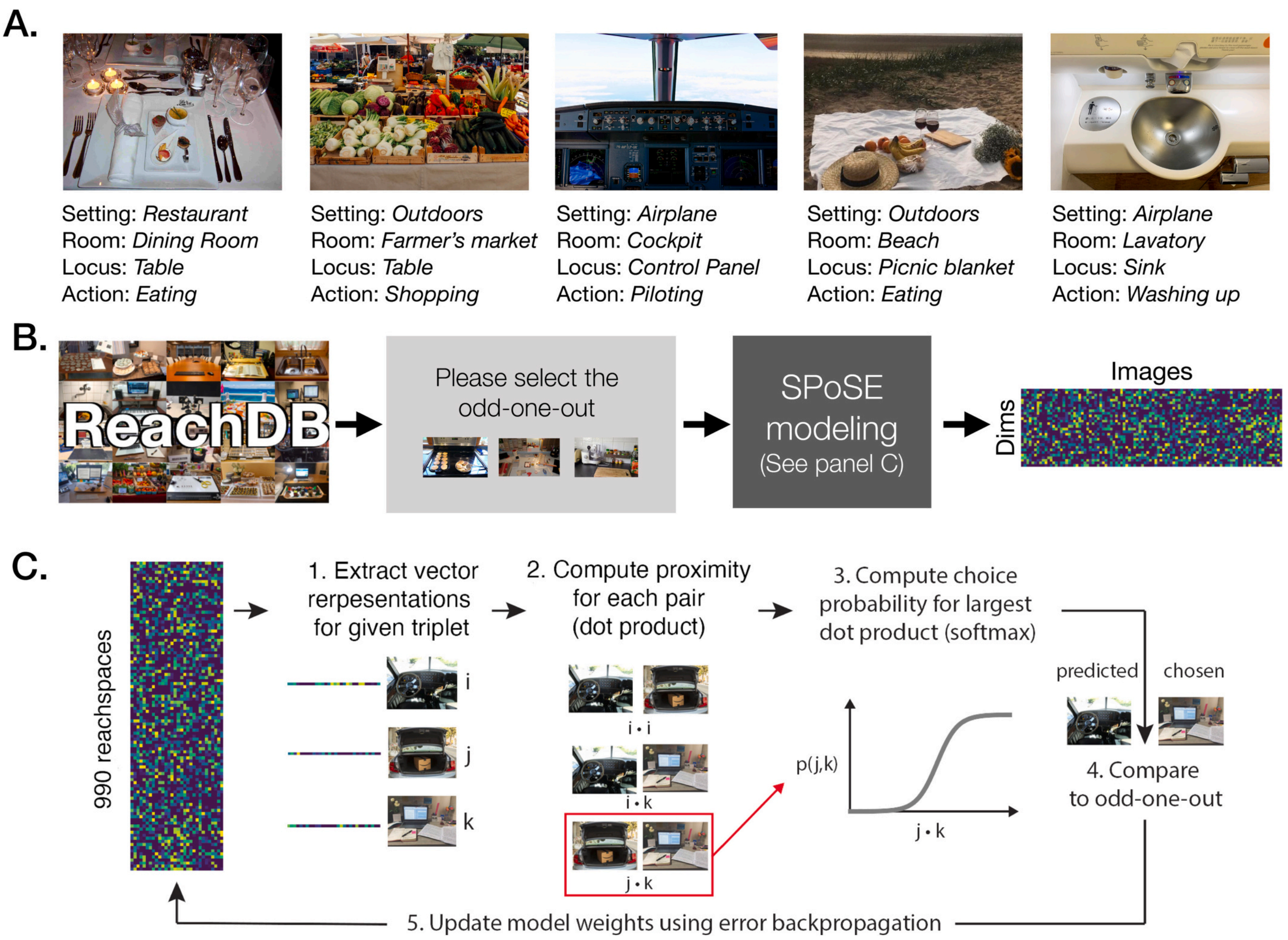


Fig. 1. Stimuli and Methods. A) Examples of reachspaces: reach-relevant environments that support task-oriented behavior using the hands. Below each image, we include the associated labels from the Reachspace Database, which annotate the location of the reachspace at three levels (setting, room, interaction locus), and the action it supports. B) Steps of the modeling procedure: 990 images were selected from the Reachspace Database (ReachDB), with 3 exemplars from each of 330 categories. Odd-one-out judgments were collections on 1.25 million triplets, and modeled using the Sparse Positive Similarity Embedding approach (Hebart et al., 2020), to derive a low-dimensional embedding that predicted human similarity judgments. C) Detailed illustration of the SPoSE modeling approach (reproduced and modified with permission from Hebart et al., 2020)

similarity judgments among pairs of reachspaces are predicted more strongly by their respective functions than their locations. Altogether, this work reveals the structure of internal representations of reach-relevant spaces and highlights the broader importance of function for organizing knowledge about the world.

2. Methods

2.1. Participants

A total of 4269 Amazon Mechanical Turk (MTurk) workers were recruited for this work (mean age was 37.78, 45.2% female, 0.3% non-binary). All workers were based in the US and had MTurk performance approval ratings over 90%, with a minimum of 500 HITs (i.e. "Human Intelligence Task", the term MTurk uses for a task assignment) completed. Workers gave informed consent and were compensated for their participation. These participants were divided among several separate tasks: 3075 Workers provided odd-one-out judgments for the main task, 322 Workers provided odd-one-out judgments for a 45-image validation task, and 376 Workers provided odd-one-out judgments for the reliability sample. An additional 447 Workers participated in the image rating task, and 49 Workers provided labels for dimensions in a naming task. All procedures were approved by the Harvard University Human Subjects Institutional Review Board.

2.2. Stimuli

The stimulus set consisted of 990 images of reach-relevant environments, selected from a beta version of the Reachspace Database. (Josephs, Zhao, & Konkle, 2021) There were 330 reachspace categories, with 3 images per category, chosen to be as different from each other as possible while remaining good examples of the category. All images depicted near-scale views (between 2 and 4 ft in distance, i.e. the distance one typically sits from a desk) of environments that afford hand-based actions, captured from the point of view of an agent performing a task in the space (See Fig. 1A for examples). The depicted reachspaces generally consisted of extended surfaces, oriented horizontally or vertically, populated with objects, buttons or other elements that afford interaction. Category labels in the Reachspace Database are composed of four taxonomic levels, which index the following: the setting it belongs to (e.g. hotel, home, office building, the outdoors), the room or site it occupies (e.g. dining room, conference room, campsite), the primary structure supporting the interaction with the environment ("interaction locus", e.g. surfaces such as tables and shelves, or large interactable objects like control panels and digital kiosks), and the specific action it affords (e.g. cake decorating, titrating). A given category in the database corresponds to a specific combination of these levels. Category labels were determined manually by the creators of the database based on the search terms used to discover the images and validated by an additional lab member without personal stake in the project. Image display size could vary according to individual Worker computer parameters, but images always maintained an aspect ratio of 4:3, and the maximum display size was 400 × 300 pixels for the triplet odd-one-out task, and 200 × 150 in the other tasks.

2.3. Behavioral tasks

Triplet odd-one-out task: Similarity judgments among reachspaces was measured in an "odd-one-out" task. In a given trial, participants were shown three images side by side and asked to indicate with a click which image was the odd one out. Images remained on the screen until the participants made their response, and trials were untimed. At the beginning of the task, participants were told to "imagine yourself in the environments: where are you standing, what are you holding, what are you doing?", but given no additional guidance.

A single triplet judgment yields information about three pairs of objects: it indicates that the selected image has low similarity with each of the two non-selected images, and the non-selected images are similar to each other. In this approach, the third image acts as a minimal context within which to evaluate the similarity of the other two images. Across many trials, a given pair of images is thus evaluated across different contexts, allowing us to measure the probability that two images will be considered similar, across a range of different contexts. Images triplets were randomly assembled from the stimulus set, with the constraint that every possible pair of images was sampled at least once (median = 10, range = [1, 28]).

This task was conducted in sets of 20 trials, and Workers could perform as many sets as they wanted up to 250 sets (mean = 19.9, range = [1, 250]). Image triplets were randomly assigned to trial sets. These judgments were used as the basis of a reachspace embedding (see Modeling section).

Image rating task: To assign labels to clusters discovered in the reachspace embedding, we collected correspondence ratings for all images on experimenter-generated labels. For the rating task, participants were given a description, such as "For this task, please indicate which images are related to electronic equipment: that is, images that are related to electronics, computers, and other digital equipment", and indicated which images corresponded to the label (see Supplement all instruction wordings used in this task). Participants received only one label per task and rated all images against the same label. Each trial consisted of a five by five array of images, and participants clicked to select images fitting the label. Selected images were highlighted with a red border, and could be unselected by clicking again. To prevent the task from being too long, the image set was divided across three separate task sets, and participants could perform as many sets as they wanted. Each image was seen only once per set, with the exception of 20 duplicate images which were used for quality assurance (subjects with <75% agreement on these duplicate images were excluded from analysis).

Dimension naming: Common-sense labels were obtained for each of the 30 dimensions comprising the embedding in a simple naming task. Participants saw a 4-by-3 array of reachspace images and were asked to name the property displayed in the images. Each array consisted of images selected from the top of one dimension from the embedding. Arrays were created by randomly selecting 12 of the top 20 images for that dimension. To ensure that dimension labels were not influenced by the exact images included, 5 such arrays were created per dimension, yielding 5 different random samples of 12. A given participant saw only one array for a given dimension. Participants were asked to type up to 5 possible labels that described the images in the array, keeping them to 1–2 words in length. Dimensions were presented in random order to minimize order effects. There were 32 trials, one for each dimension, and a compliance-assessment trial consisting of a 4-by-3 array of beach scenes.

2.4. SPoSE computational modeling

Data from the odd-one-out task were trimmed to enforce quality. Simulations on 100,000 trials showed that stricter data trimming criteria led to more dimensions, higher accuracy, and lower final loss (i.e. cleaner data was better than more data), so data were cleaned with the following strict criteria before modeling and analysis: 1) all individuals with 60 trials or more who used the same key for >40% of HITs were removed, 2) all HITs where participants responded with more than three consecutive sequences (e.g. position 1 then 2 then 3, repeated 3 times) were removed. To set these criteria, we performed simulations of what positions the odd-one-out would occupy in the triplet displays given that image positions were randomly selected, in the hypothetical case where the odd-one-out is known. We found that the odds of the odd-one-out falling in the same position in >40% of the trials, were <10%, and the probability of more than three consecutive sequences were <5%. Finally, all trials with reaction times >3 standard deviations from the mean were removed (reactions times were log transformed prior to trimming to account for the right-skew of reaction time distributions). In total, 1,251,823 trials passed quality assurance and were used in the modeling (out of 1,716,697 collected).

An embedding for the images was derived from these data using the Sparse Positive Simiarity Embedding approach.(Hebart et al., 2020) The embedding is expressed as a matrix with images in the rows, and dimensions in the columns, where the value of each cell expresses how strongly a given dimension is present in a given image. Details of the modeling procedure are given in Hebart et al. 2020, but briefly, an embedding is initialized with random weights (range 0,1) on 90 latent dimensions for each of the images, yielding a 990-by-90 matrix. The embedding is then tuned using stochastic gradient descent trained on the human odd-one-out responses. The objective function being minimized consists of a cross-entropy loss between the predicted triplet choice probabilities for all three choice options and the actual choices, added to an L1-norm on the weights to enforce sparsity. The trade-off between these two terms is estimated using the regularization parameter lambda ($\lambda = 0.007$), which we tuned in a pre-analysis hyperparameter-tuning step performed on approximately 80% of the data and which we optmized to yield the lowest final loss (the same procedure was used to tune the learning rate, lr = 0.0005). Prior to training, the dataset is partitioned into a training set and testing set (90/10 split), then training proceeds by making odd-one-out predictions based on the embedding, and evaluating performance on the withheld testing set after every epoch (max 500 epochs), until the model converges (model convergence criterion: no change in training accuracy over the last 50 epochs, at a threshold of $p > 0.1$ using linear regression.). Finally, only the most informative subset of the 90 discovered dimensions were selected for analysis: dimensions with all weights below 0.1 were removed, yielding the final embedding.

This model procedure is stochastic, so different initializations will yield embeddings with slightly different numbers of dimensions, covering slightly different attributes. To select the most generalizable embedding, we ran 50 random initializations of the model and selected the embedding with the highest average correlation to all other embeddings. For a given pair of embeddings, their correlation to each other was determined using a split-half cross-validated analysis. Since dimensions are not labeled, the first step is to establish which dimensions in the two embeddings correspond to each other. Using half the data (i.e. weights on all dimensions for half the images), each dimension in one embedding was correlated with each dimension in the other embedding. The pair of dimensions with the highest correlations were interpreted as corresponding to each other across embeddings. Next, the remaining data was used to assess the magnitude of the correspondence, as the average correlation among all pairs of corresponding dimensions. This procedure was performed for every possible pair of embeddings, and the embedding with the highest average correlation to all other embeddings was selected for further analysis.

2.5. Analysis of the embedding

Embedding replicability The replicability of a given dimension was expressed as its replicability score, calculated as its average correlation across the 50 SPoSE iterations. For each dimension in the final embedding, the closest match was identified in each of the embeddings from the remaining 49 iterations using Pearson correlation. The average value of this correlation across all embeddings was taken as the replicability of the dimension. Names for each of the dimensions were derived in 2 ways. First, we provide concise labels corresponding to the concept we felt was most clearly illustrated in each dimension. For increased objectivity, we additionally solicited dimensions names using the behavioral procedure described above. Dimension names were collected from 50 participants, then aggregated. All labels appearing >3 times were retained and displayed using word clouds.

Noise Ceiling An additional behavioral dataset was collected to estimate the noise ceiling for the behavioral responses. The noise ceiling was operationalized as the average participant agreement on a given answer for a given triplet. One thousand random triplets were selected, and an average of 29 ratings were collected per triplet (range: 18 to 36). Auxiliary analyses confirmed that 1000 triplets were sufficient to obtain stable noise ceiling estimates. The consistency of responses for each triplet was estimated as the proportion of the time that the most popular response was chosen (100% = consistent agreement about the odd-one-out, 33% = chance). This value was averaged across all triplets. The noise ceiling was calculated as the mean consistency across all 1000 triplets. The percent of explained variance was computed as follows: (model performance – chance)/(human agreement - chance). In total, 28,177 odd-one-out trials were included for noise ceiling estimation.

Embedding Validation The ability of the embedding to predict human representational similarity was validated by randomly selecting 45 images from the stimulus set and obtaining a separate behavioral sample of odd-one-out judgments on all possible triplets for these images (26,588 total trials). A behavioral representational similarity matrix (RSM) was derived by scoring the three possible pairs of images in each triplet: pairs involving the odd-one-out were given a “0” (“considered different”), and the remaining pair was given a “1” (“considered similar”). These scores were aggregated across trials, and divided by the total number of trials per cell to yield a matrix of the proportion of times each pair was treated as similar. A model representational similarity matrix for the 45 images was derived in a similar way: we first used the embedding to predict trial outcomes for every possible triplet combination, then computed an RSM from these results using the same approach as for the behavioral RSM. The correspondence between these matrices was assessed by taking the Pearson correlation between them.

Clustering analysis: Divisions were identified among the images using k-means clustering over the embedding. We tested cluster numbers ranging from 2 to 8. For each clustering solution, the average correlation among all cluster centers was obtained. The optimal number of clusters was determined to be that at which this value starts to plateau (Supplemental Fig. 2). This method of cluster number selection yields the highest number of clusters where cluster centers show more than minimal distinctions from each other. Possible labels for each cluster were derived by looking at the images in the cluster and naming the concepts that they appeared to correspond to. The validity of these labels was assessed using the image rating task described above. Worker responses were turned into binary scores: all images selected for a given label >50% of the time was considered to match the labels, all remaining images were not. Finally, the match between each concept and the k-means clusters was assessed by taking the Adjusted Rand Index (ARI) between the vector of cluster assignments and the vector of label assignments for a given concept. The ARI measures the agreement between two clustering solutions, by finding the proportion of item pairs that are in the same cluster in both solutions plus the proportion that are in different clusters in both solutions. This value ranges from 0 (chance) to 1 (the clustering solutions are identical).

Assessing the roles of function vs *location in predicting similarity:* We tested whether similarity judgments on reachspaces reflected shared setting, room, interaction locus or action. The setting, room, locus and action for each reachspace was drawn from the category name given to each image in the Reachspace Database. Prior to the analysis, author EJ confirmed that the labels matched the images. The contribution of each of these factors was assessed with a similarity prediction score: for each of 10,000 draws, a randomly selected reference image was compared to two other images: one sharing a label with the reference (according to the given factor) and the other having a different label. Apart from the constraint imposed by the labels, comparison images were randomly selected. The similarity between the reference and each comparison was assessed using the Euclidean distance between the images' embedding weights, and the similarity prediction score was calculated as the proportion of times that the image which shared a label with the reference had the higher similarity.

RSA regression analysis: A matrix of pairwise dissimilarities was obtained for the images by measuring their Euclidean distance in embedding space. Matrices representing the pairwise overlap in labels between images were generated for each taxonomic level separately, by coding a match in the labels as 1, and a non-match as 0. The matrices were reduced to their upper diagonals and vectorized, then entered into an Ordinary Least Squares linear regression with the binary matrices for Action, Setting, Room and Locus as predictors, and the matrix of Euclidean distances as the prediction target.

3. Results

To identify attributes that shape our internal representations of the reachable world, we modeled the similarity space of a broad sample of reachspace images, based on human similarity judgments. A set of 990 reachspace images, spanning a diverse collection of 330 different categories, were selected from the Reachspace Database.(Josephs et al., 2021) All images depicted near-scale views of spaces that afford hand-based actions, captured from the point of view of an agent performing a task in the space (See Fig. 1A for examples). The depicted reachspaces generally consisted of extended surfaces, oriented horizontally or vertically, populated with objects or other elements that afford interaction, such as tabletops, countertops, or digital kiosks. Image categories were selected to widely sample the reachable world, including places where we work, play, study, eat, shop, create, perform music, store things, and more. These categories are highly granular, dividing up reachspace types according to a combination of the settings they are found in (e.g. home, hotel), the rooms they belong to (e.g. office, dining room), the locus of interaction in the space (e.g. desk, counter), and specific actions associated with the depicted reachspace (e.g. working, eating, cake decorating).

A similarity space was derived from these judgments using the Sparse Positive Similarity Embeddings approach, or SPoSE,(Hebart et al., 2020) which begins by randomly initializing each image as a point in a high-dimensional feature space, and then tuning image weights along these dimensions to derive an embedding which predicts behavioral judgments from the triplet task. Notably, this approach yields a dimensional model of the similarity space, inferring a set of axes which underlie the variation among images and assigning each image a score on each of these dimensions. This model makes two theoretical assumptions: first, it assumes that the dimensions of this embedding space are sparse — that is, that each reachspace only has weights on some dimensions, but not all (e.g. a labbench would have low weights on dimensions relating to food or leisure). Second, it assumes that the dimensions are positive, such that they can only add up but not cancel out (e.g. a food-related property and a seating-related property should not cancel each other out). This also means that the weight of an image on a dimension can be interpreted as the amount of the corresponding property present in the image. These twin assumptions regarding the dimensions are in contrast to other common dimensionality reduction methods such as principal component analysis (PCA), which assume dense dimensions spanning the entire range of real numbers. Additionally, SPoSE allows dimensions to be cross-correlated, while PCA assumes uncorrelated dimensions. As a result, larger numbers of dimensions tend to be discovered with SPoSE than with PCA, and they tend to correspond to finer-grained details (e.g object parts or properties,(Hoyer, 2002; Hoyer, 2004)) and the tend to be more interpretable than PCA dimensions. With SPoSE, the weight of an image on a dimension can be interpreted as the amount of the corresponding property present in the image.

3.1. Model validation

First, we established that the modeling approach yielded a stable, accurate, and replicable model of human similarity judgments. As the SPoSE modeling approach is stochastic, we ran 50 iterations, yielding 50 embeddings. Solutions were largely stable: after dropping dimensions with low weights (any dim with max weight < 0.01, see Methods) the number of discovered dimensions ranged from 27 to 32, and model accuracy on the test set had very low variance across the 50 iterations (mean: 60.55%, stdev: 0.04%).

We next selected one embedding on which to perform our analyses. We identified the embedding that was most representative of all 50 SPoSE iterations, operationalized as the embedding with the highest average correlation with the 49 other embeddings (see Methods). This model contained 30 dimensions. Finally, we confirmed that the dimensions in the selected embedding were replicable, that is, that they consistently appear in different iterations of the model. A replicability score was calculated for each dimension by correlating it with the corresponding dimension in all other embeddings (see Methods). All dimensions had replicability scores greater than $r = 0.70$ (maximum possible value: 1), with the exception of one dimension corresponding to outdoor arrays of objects, whose score was $r = 0.40$. In total, 20/30 dimensions had replicability scores >0.90, and 26/30 had >0.80. This suggests that the results we report are not specific to the individual model iteration we examine, but instead are stable properties of the embedding space for these images.

Next, we examined whether the model successfully learned to predict human odd-one-out judgments. After training, the model achieved 60.6% accuracy on a held-out set of triplet judgments (Fig. 2, chance = 33.3%). To contextualize this performance, we estimated the noise ceiling of the behavioral data: in a separate behavioral sample, odd-one-out judgments were collected for 1000 triplets, randomly sampled from the image set. The average consistency in judgments across participants was 66.3%. Thus, the SPoSE model could predict human behavior up to 82.1% of the noise ceiling (see Methods).

The SPOSE approach infers an image embedding from a fraction of the total possible triplet judgments, which reduces data collection to tractable levels (1.25 million trials represents about 1% of the total number of triplets possible with 900 images). Does the embedding yielded by the model accurately reflect the embedding we should expect if it were derived purely from behavioral judgments in the hypothetical case where we could fully-sample all triplets? We estimated this by collecting all possible triplet judgments for a subset of 45 images, deriving an embedding for these images purely from the behavioral data, then comparing it to the embedding yielded by the model (see Methods). Overall, behavioral and model RSMs were highly correlated ($r = 0.95$, Fig. 2C), validating that the embedding we analyze closely reflects the similarity space underlying human behavioral judgments.

Overall, the SPOSE approach yielded a stable and replicable model that was highly successful both at predicting triplet similarity judgments and at reconstructing the representational space underlying these judgments. This suggests that human judgments of reach-relevant environments are principled and structured. In the following sections, we examine this embedding to reveal key principles that underlie these judgments.

3.2. The embedding dimensions are interpretable and informative

A significant benefit of the SPoSE modeling approach is that it yields an embedding with accessible and interpretable dimensions. We first examined these dimensions, to gain insights about the properties that are salient to observers when making similarity judgments on reachable environments.

Each dimension was visualized by ordering the images according to their weights on the dimension (Figs. 3, 4, and Supplemental Fig. 1 show the top 6 images per dimension, with word clouds depicting participant-generated labels for the dimension). While the dimensions emerge independently in the model, here we discuss them in pairs or groups, to better highlight some of the concepts they capture. First, some of the dimensions pertain to global properties of the space: for example, separate dimensions emerge for cluttered versus clear spaces (Fig. 3A). Additionally, many of the dimensions capture complex combinations of

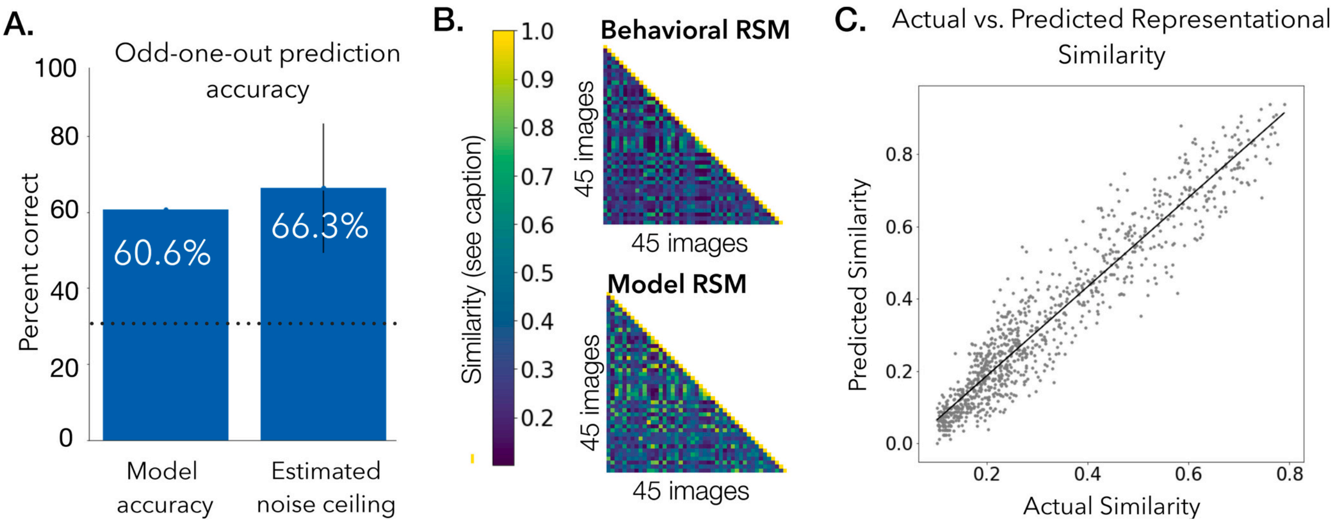


Fig. 2. Validation of the model embedding derived from similarity judgments over 990 reachspace images. A) Model performance on odd-one-out prediction for held-out test set. The noise ceiling of the behavioral data was estimated from a separate behavioral sample, and represents the average inter-rater reliability over 1000 triplets. B) Representational similarity matrices for a 45-image subset of the stimulus set, created by fully sampling all possible triplets in a validation behavioral experiment (top) and by estimating similarity based on the model embedding (bottom). Here, the similarity between two images is operationalized as the proportion of times they are judged to be similar, across all trials. C) Correlation between actual and predicted similarity between all image pairs in (B).

semantic category and physical affordance information about a space. For example, two dimensions emerge for musical instruments, distinguishing between keyed instruments and non-keyed instruments (Fig. 3B). Likewise, two dimensions emerge for outdoor spaces, distinguishing those containing multiple small objects from single large objects (Fig. 3C), and separate dimensions emerge for workshop-related spaces where the primary surfaces was oriented vertically vs horizontally (Fig. 3D). Third, some dimensions also captured information about the intended user of the space: separate dimensions emerge for children's games versus adults' games (i.e. gambling, Fig. 4A) and for electronic spaces used by everyday consumers versus those requiring expertise (Fig. 4B). Fourth, some dimensions capture information about the physical properties of the space: craft-related spaces have separate dimensions for arts which use wood vs other media (Fig. 4C) and other dimensions emerge for spaces with ceramic, paper, or stainless steel components (Supplemental Fig. 1). Additionally, multiple dimensions emerge relating to the storage of items, with distinctions between storage at home, in retail, with portability constraints, and for travel (Supplemental Fig. 1; see panels 3E and 4D for additional dimensions not discussed here).

Note that any given dimension captures a complex combination of attributes, and thus can be characterized in multiple ways. Here we discuss just one interpretation per dimension to highlight the kinds of concepts they measure. Overall, dimensions discovered by the model were interpretable and revealed fine-grained distinctions within this large set of reachable environments.

3.3. Evidence for conceptually distinct reachspace classes

We next characterized the global structure of reachspace similarity in this 30-dimensional space. Similar work on objects,(Hebart et al., 2020) showed global organization into well-known classes (animate/inanimate, natural/human-made). The current study drew from 330 different reachspace categories, but it is possible that these are conceptually clustered into a smaller number of classes. How many classes were participants sensitive to in these images, and what concepts do they correspond to?

Taking a data-driven approach, we applied k-means clustering to group the images according to their similarity in the embedding space (see Methods). Fig. 5A shows a 2D projection of the representational space for the 990 images, with cluster assignment indicated by color (the projection was obtained using MDS-initialized t-SNE, to capture both global and local structure). Visual inspection of the images in each cluster (Fig. 5B) suggests that they relate to 1) food and eating, 2) computers and electronics, 3) spaces for storage, retail and display (excluding food), and 4) entertainment, hobbies, and handicrafts. The fifth cluster is less interpretable and contains a mix of spaces related to drinks, miscellaneous liquids, and household chores. Thus, human similarity judgments suggest the existence of about 4–5 broad types of reachspaces within our sample.

To validate these possible cluster identities, we collected behavioral ratings for the images on Mturk and assessed the correspondence between ratings and cluster assignments (see Methods). For each possible cluster identity described above, participants were presented with a brief description (see Supplemental Table 1 for task wording) and asked to indicate for each image whether it matched the concept or not. Correspondence between these conceptual labels and the k-means clusters was assessed using the Adjusted Rand Index (ARI), which calculates the proportion of times both solutions agreed about whether a pair of images was in the same cluster or not (ARI = 0 indicates chance, ARI = 1 indicates perfect agreement between the solutions). Results are shown in Fig. 5C and D. A correspondence matrix between cluster identity (cluster 1–5) and conceptual label confirmed that each of the hypothesized cluster identities accounte for different clusters (Fig. 5C). The clearest clusters corresponde to food-related and electronics-related reachspaces, with near perfect alignment between images with those attributes and clusters 1 and 2 (ARI = 0.80 and 0.76, respectively). Cluster 3 alignes moderately-well with retail spaces (ARI = 0.67) and with the broader concept of spaces designed for storage and display (ARI = 0.44). Cluster 4 correspondes moderately-well with spaces for games, hobbies, art and handicrafts (ARI = 0.36). The final cluster is less interpretable and shows low but above-chance correspondence to drinks- and liquids-related reachspaces (ARI = 0.22) as well as spaces related to household chores (ARI = 0.22). These correspondences between clusters and labels can also be visually assessed by comparing the clustering in Fig. 5A with the visualization of image labels in Fig. 5D. Overall, this analysis shows that the 900 reachspaces in our sample can be divided into a relatively small number of broadly distinct classes. These classes may point to important global division in the space of reachspaces, similar to animate/inanimate and indoor/outdoor distinctions that divide objects

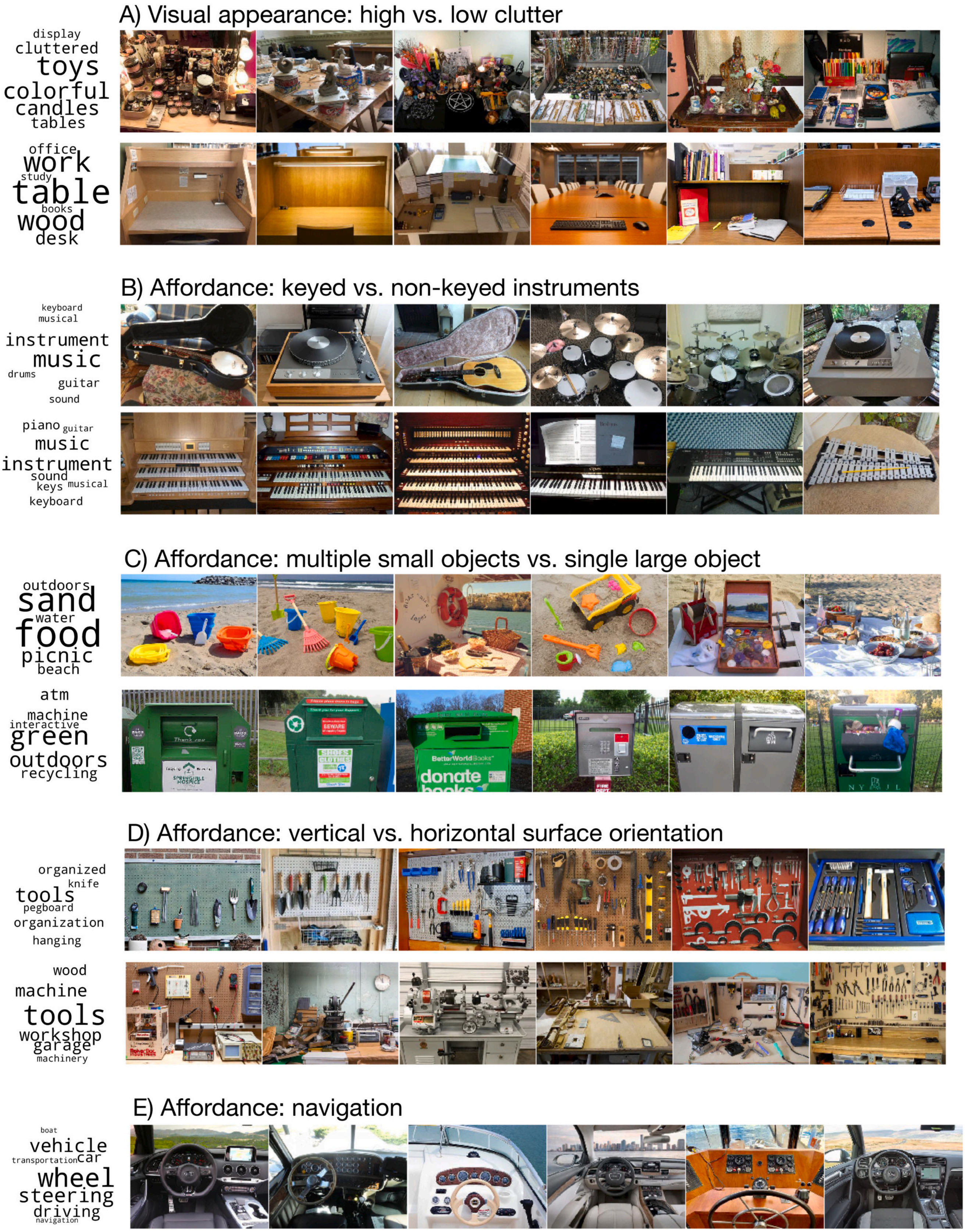


Fig. 3. With Fig. 4 and Supplementary Fig. 1, this figure shows some of the dimensions forming the embedding. Each dimension is illustrated with the top 5 images on the dimension, and along with a word cloud which shows responses from 50 participants asked to judge what is captured by the dimension (bigger text = more common label). Here, we have divided the dimensions into groups for the purpose of illustrating some of subtle distinctions they are sensitive to, however note that this is just one way of considering them.

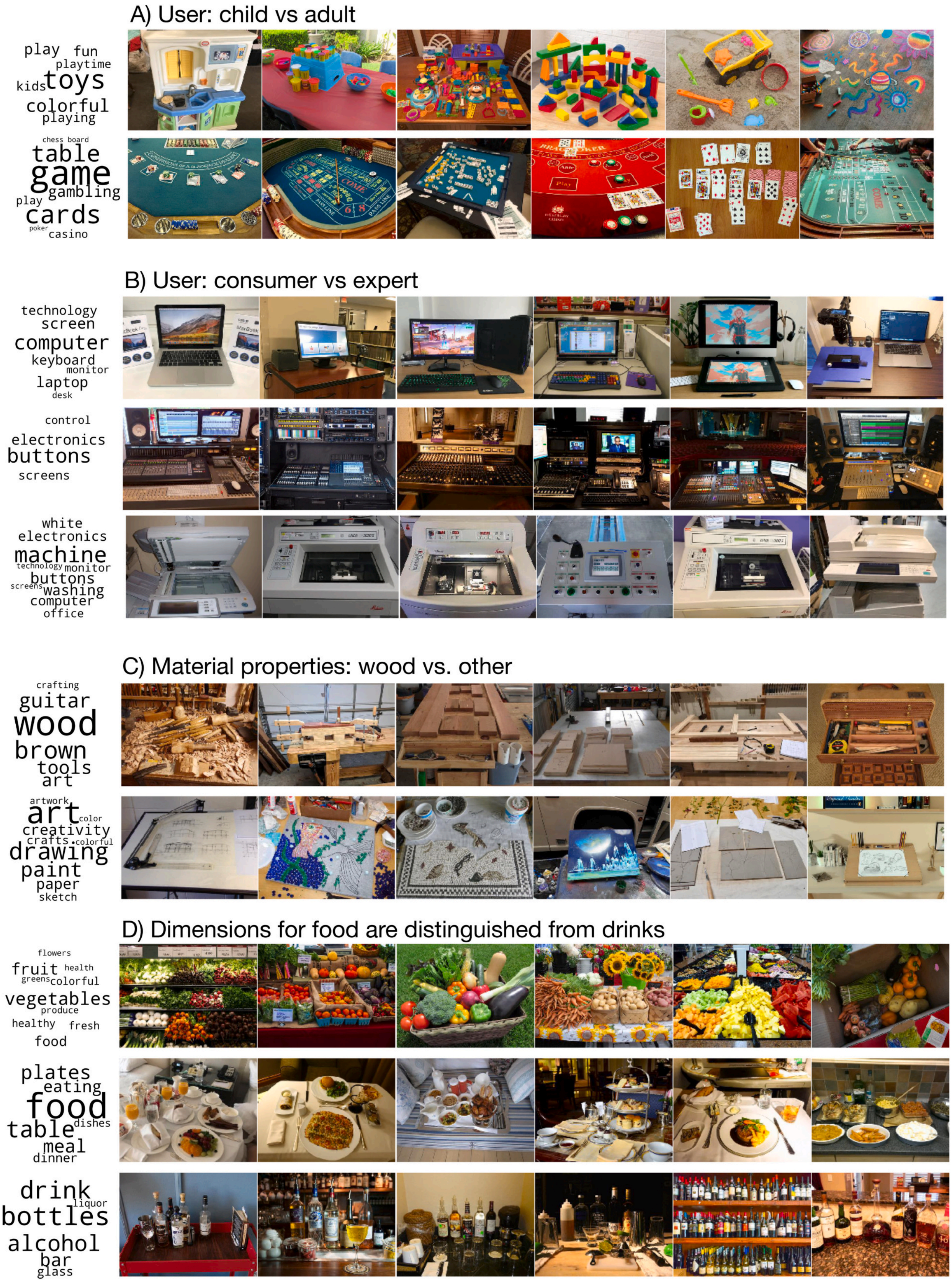
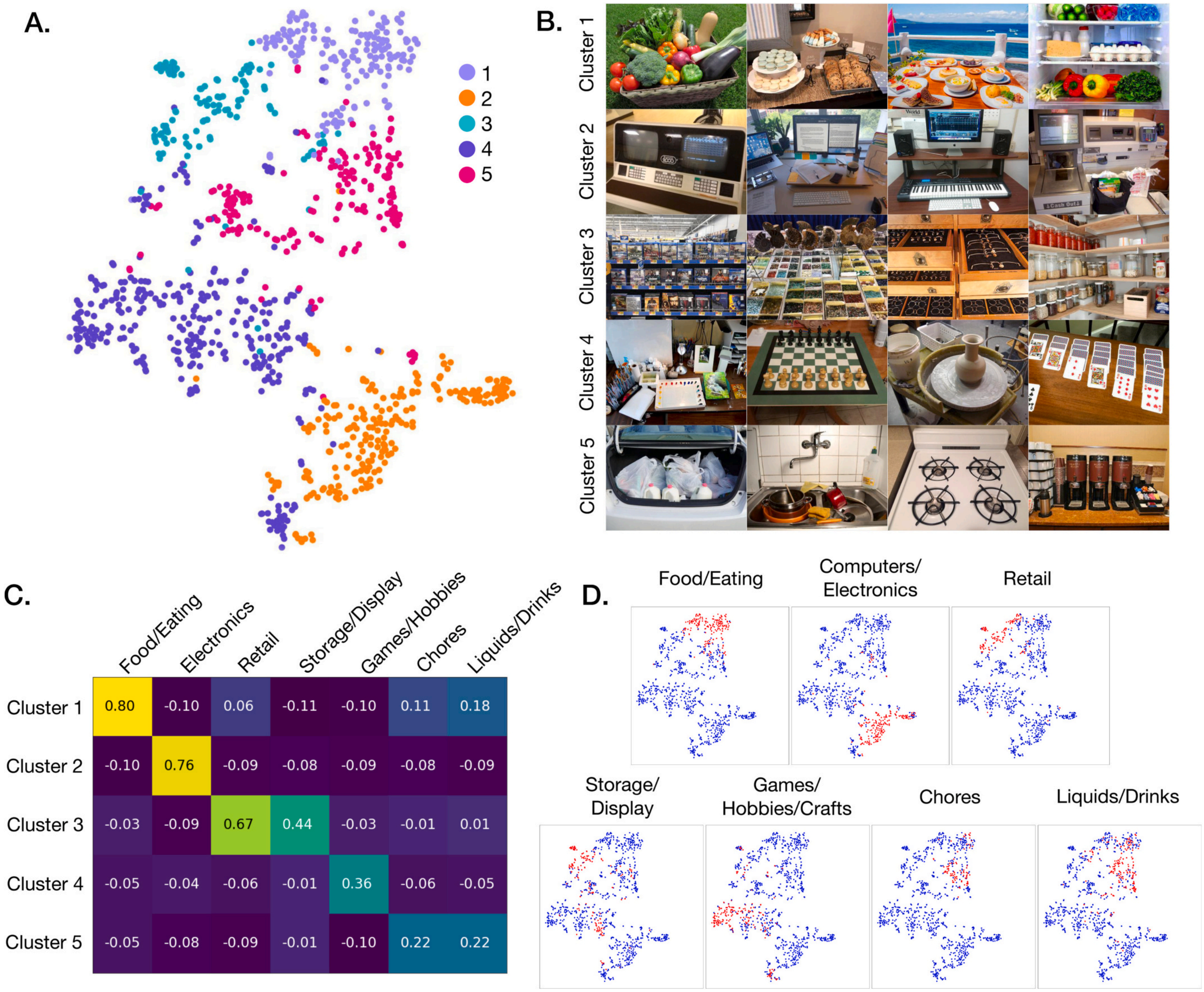

A) User: child vs adult
play fun playtime kids toys colorful playing
chess board table game gambling play cards poker casino
B) User: consumer vs expert
technology screen computer keyboard monitor laptop desk
control electronics buttons screens
white electronics machine technology monitor buttons screens washing computer office
C) Material properties: wood vs. other
crafting guitar wood brown tools art
artwork art color creativity crafts colorful drawing paint paper sketch
D) Dimensions for food are distinguished from drinks
flowers fruit health greens colorful vegetables produce healthy fresh food
plates eating food table dishes meal dinner
drink liquor bottles alcohol bar glass

Fig. 4. With Fig. 3 and Supplementary Fig. 1, this figure shows some of the dimensions forming the embedding. Each dimension is illustrated with the top 5 images on the dimension, and along with a word cloud which shows responses from 50 participants asked to judge what is captured by the dimension (bigger text = more common label). Here, we have divided the dimensions into groups for the purpose of illustrating some of subtle distinctions they are sensitive to, however note that this is just one way of considering them.



	Food/Eating	Electronics	Retail	Storage/Display	Games/Hobbies	Chores	Liquids/Drinks
Cluster 1	0.80	-0.10	0.06	-0.11	-0.10	0.11	0.18
Cluster 2	-0.10	0.76	-0.09	-0.08	-0.09	-0.08	-0.09
Cluster 3	-0.03	-0.09	0.67	0.44	-0.03	-0.01	0.01
Cluster 4	-0.05	-0.04	-0.06	-0.01	0.36	-0.06	-0.05
Cluster 5	-0.05	-0.08	-0.09	-0.01	-0.10	0.22	0.22

Fig. 5. Data driven discovery of large-scale divisions in the representational space of reachspaces. A) 2-D projection of the representational space using MDS-initialized tSNE. Dots correspond to images, and are colored according to their clustering assignment in k-means clustering (k = 5). B) Four example images from each cluster. C) Adjusted Rand Index measuring the image-wise correspondence between cluster assignment and labels derived from behavioral ratings (0 = no better than chance, 1 = perfect correspondence). D) Visualization of the embedding space with each behaviorally-derived label shown. Red dots indicate the images that were judged to fit the given labels in a behavioral tasks. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

and scenes, respectively.

3.4. *Reachspace similarity judgments reflects shared function more than shared location*

Recent studies of navigable-scale environments found that the function of a place plays a large role in determining what other places it will be considered similar to (Greene et al., 2016). Are reach-relevant environments likewise grouped by function? On the one hand, reachspaces are designed to support specific activities, so they may show strong conceptual organization by function. However, reachspaces are related to the broader environment in lawful ways and tend to belong to particular locations (e.g. a kitchen counter is in a kitchen, a bedside table is in a bedroom). Thus, they may be better grouped according to the locations they occupy. We next tested which of these principles – function or location – best accounted for similarity judgments among reach-relevant environments.

Function and location labels for each image were obtained from the labels provided in the Reachspace Database (Josephs et al., 2021)(see Methods). The function of a reachspace was operationalized as the action it affords (similar to Greene et al., 2016), for example “shopping”, “chopping vegetables”, “cake decorating” and “embroidering”. These actions are defined at a high level of specificity (e.g. “chopping vegetables” and ”rolling dough” rather than the more general “cooking”), as this is a better reflection of the precise activity, object array, and motor plans associated with a given reachspace. The location of each reachspace was operationalized in three different ways. From the broadest to the most specific level, location was operationalized as the setting the reachspace is located in (e.g. office building, home, hotel, hospital), the room containing the reachspace (e.g. kitchen, office, bedroom), and the interaction locus (hereafter locus) of the space, i.e. the type of object or surface that forms the primary support structure (e.g. desk, table, pegboard, counter, control panel). Note that while these three labeling schemes describe context at different scales of spatial inclusion, they are not nested hierarchically (a table can be found in many different rooms, and rooms such as offices can be found across many different settings) and should instead be thought of as different, partially independent ways of slicing across the images (for an illustration of this, see the labels in Fig. 1A).

Overall, reachspaces were divided into 38 Settings, 122 Rooms, 151 Loci, and 131 Actions. These labeling schemes were highly independent from each other (average Adjusted Rand Index among the different labeling schemes was 0.14, see Supplemental Analysis 1). We quantified whether sharing a label under each of these schemes predicted greater similarity than having different labels. Over 10,000 iterations, we randomly selected one reference image and 2 comparison images, with the constraint that one comparison image shared a label with the reference and the other did not. We then measured the proportion of times the reference-comparison pair which shared a label was more similar (i.e. lower Euclidean distance) than the pair that did not. Results are shown in Fig. 6.

Sharing an Action label predictes relative similarity among images 75.1% of the time, significantly more than sharing a label at the Room level (67.8%, z score of comparison = 12.7, $p < 0.001$, one-sided Two Proportion *Z*-test), Locus level (65.1%, z = 16.6, $p < 0.0.001$) or Setting level (62.3%, z = 20.16, p < 0.0.001). However, all four labelling schemes predict greater similarity at above chance levels (Action: z = 62.6, p < 0.001; Room z = 40.7, p«0.001; Locus z = 33.5, p < 0.001; Setting z = 28.0, p < 0.001; one-sided single sample Proportion Z-test).

To evaluate the concurrent relative role of these different labelling schemes in predicting image similarity, we conducted Representation Similarity analysis (RSA) in the form of RSA regression. For each labeling scheme, we generated a binary matrix capturing whether each pair of images shared a label. These matrices were entered into the regression as predictors, and the pairwise Euclidean distance among images in the SPoSE embedding was the prediction target (see Methods). We found that each of the labelling schemes independently account for some of the variance in image dissimilarity, confirming the above results (p < 0.001 for each of the regression coefficients). Additionally, by examining the regression coefficients, we found that Action accounts for the most variance (−0.37, note that coefficients are negative because sharing a label predicts smaller Euclidean distances), followed by Room and Locus (which were roughly equivalent at −0.224 and − 0.222 respectively), and finally the Setting level (−0.11).

Thus, both the location and the afforded action account for some of the structure in the representational space of reachspaces, but action is a better predictor of representational similarity. Overall, this suggests that human judgments of similarity among reachspaces relate more to the function they serve than the places they occupy.

4. Discussion

Here, we used 1.25 million human similarity judgments to derive an embedding for 990 images of reach-relevant environments, and examined this embedding to characterize key factors that organize our conceptual representations of the reachable world. We found that human similarity judgments can be predicted with a 30-dimensional representational space and that the embedding dimensions capture information relating to the content, layout, purpose, and even typical user of the space (e.g. adult vs child). We described the global structure of this similarity space, finding evidence for a small number of broad reachspace classes, and we found that function is a better determinant of conceptual similarity than location. Altogether, this work reveals the conceptual structure of reach-relevant environments, in a large-scale, data-driven manner.

4.1. Dimensions of reachspace similarity space

What kind of information is captured by the dimensions? In general, dimensions discovered by the model appeared to capture multiple attributes (e.g. "*vertically* oriented *storage* in a *workshop*" in Fig. 4D). This reflects the consistent finding that attributes in the world are "clumped" and covary with each other, rather than being uniformly distributed. (Rosch et al., 1976) Additionally, the discovered dimensions capture both low-level visual information and high-level semantic information (e.g. the dimension for child-related spaces also featured bright colors and mid to high levels of clutter). For interpretability, we have discussed the dimensions using relatively high-level labels, but it is possible that similarity judgments rely equally on the covarying lower-level, perceptual features. Indeed, scene perception research suggests that the low- and mid-level visual appearance of an environment is diagnostic of, and to some extent inseparable from, its higher-level identity and function.(Groen, Silson, & Baker, 2017) Future work is needed to establish the content of the individual dimensions discovered here and understand their relation to the locations, functions, material properties, visual appearance (and more) of the reachspaces they describe.

It is important to note that the precise content of the dimensions is shaped by methodological and stimulus choices. Our aim was to capture a general, intuitive similarity space, so we used a triplet task with minimal instructions. Different similarity spaces would emerge if participants were asked to judge similarity on specific bases. It is also an open question how embedding spaces derived from image-computable features would compare to the behaviorally-grounded embedding we describe here. It is possible that triplet networks(Hoffer & Ailon, 2015) trained on the same images, without a human's knowledge of semantics or affordances, would yield embeddings with different focus. Likewise, it is possible that a model which tries to account for both visual and semantic similarity (for example through the addition of a loss function predicated on image similarity prediction) would yield finer-grained clusters. Additionally, we note that the number of dimensions

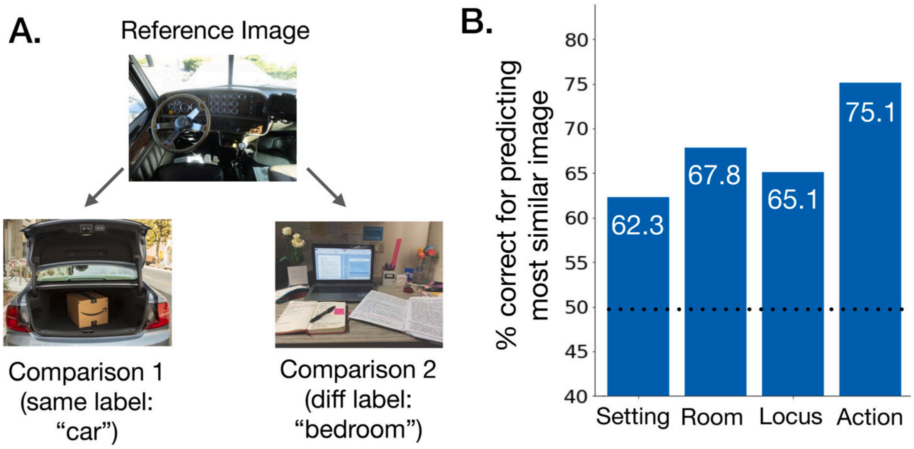


Fig. 6. Evaluating the relative influence of location and function on reachspace similarity judgments. A. Method: over 10,000 iterations, we randomly selected one reference image and 2 comparison images, with the constraint that one comparison image shared a label with the reference and the other did not. We then measured the proportion of times the reference-comparison pair which shared a label was more similar than the pair that did not. Four sets of labels were used, indexing the Setting, Room, or Locus the reachspace belongs to and the Action it supports. B. Results: Bars indicate the percent of time that the reference image had higher similarity to the comparison image which shared a label. The dotted line indicates chance (50%).

discovered depends in part on the stimulus set. If some areas of the stimulus space were oversampled, this could lead to an inflated number of dimensions spanning this part of the space, and conversely the model could not learn dimensions for areas of the stimulus space that are undersampled. One constraint of the SPOSE method is that it cannot at present be used to make predictions about images and classes that were not included in the training set. However, in spite of this, several design choices increase the chance that the results will generalize to other samples: the sparsity constraints on the model encourages it to discard spurious dimensions, and the set of reachspaces was carefully sampled to constitute a comprehensive set, both within and across categories.

How do the dimensions underlying reachspace similarity judgments compare to those for objects? The SPoSE approach was initially applied to object images,(Hebart et al., 2020) so we can compare the embeddings from the two studies. Comparing the dimensions from our Figs. 3,4, and Supplemental Fig. 1 to the dimensions in Extended Data Fig. 2 from Hebart et al. (2020), we find that some of the dimensions found here have some correspondence with dimensions for objects, most notably those for electronics and food. However, there are some major differences. First, object dimensions in Hebart et al. (2020) showed more reliance on simple features like color (e.g. black, red) or shape (e.g. round, disc-like, long and thin). In contrast, dimensions for reachspaces show evidence of integrating over more complex feature combinations, as discussed above. Second, some dimensions that appeared for objects are enriched with contextual information for reachspaces. For example: objects have one dimensions for “clothes”, but reachspace dimensions distinguish whether the clothes are in an environment relating to retail, storage, or travel (Fig. 5). Overall, while there is some overlap in the relevant concepts, the representational space of reachspaces cannot be reduced to that of individual objects.

It is more difficult to assess the generalization to scenes, as the SPoSE approach has not yet been applied to scenes. Previous work extracting scene attributes from text descriptions (Patterson et al., 2014) found that scenes properties relate to their functions (e.g. sports), prominent materials (e.g sand, foliage), material properties (e.g. rusty, glossy), and spatial envelope (e.g. open, enclosed). However, no dimensionality reduction was applied to those results, so the attribute list is large (102 attributes) and does not capture the correlation structure among them. Going forward, it will be important to test objects, reachspaces, and scenes in the same paradigms to discover the attributes that are common across them and those that are specific to different scales of experience.

4.2. Major classes of reachspaces

A clustering analysis discovered four identifiable classes of reachspaces, corresponding to food-related spaces, electronics related spaces, hobby/craft/entertainment-related spaces, and storage/retail/display spaces. While these labels provide a description of the reachspace categories in each cluster, we suggest the following broader interpretation of the classes: 1) food-related spaces, 2) digital spaces, 3) active analog spaces intended for functional engagement with objects, and 4) passive analog spaces intended for the storage or display of objects. There was an additional cluster which was ambiguous, showing weak correspondence to both household chores or drinks/liquids, but due to its ambiguity, we do not provide a broader interpretation for it.

Why might our internal representations of reachspaces show global divisions between food-related, electronic, active analog and passive analog spaces? One possibility is that human agents interact with each of these spaces in generally different manners. Acting in analog spaces usually requires manipulating physical objects in the performance of a task, and requires reasoning about the location and relations of objects across time. Active vs passive analog spaces require different amounts of interaction and monitoring over time, and food-related spaces involve additional reasoning about physiological states like hunger or appetite. In contrast, in electronic spaces, events are largely invisible and instantaneous, without physical grounding, and agents must act on simple symbols (e.g. cursors, buttons), whose function are given by learned input-output mappings. Some differences also exist in the components of environments from the different classes. Analog spaces have objects that can be moved or manipulated independently, while electronic spaces often feature components that are attached to a main structure, such as buttons, keys and switches. Thus, these four classes may reflect differences in the representations required to behave in different environments. There are many possible differences between food-related, electronic, active analog and passive analog spaces (e.g. their affordances, the amount of training and expertise they require, how common they are in daily life, the materials they are made of, their visual appearance, etc), and future experiments will be needed to establish which properties are most predictive of class boundaries.

4.3. Function as a major determiner of intuitive similarity

One major implication of this work is that function is a salient factor organizing our knowledge of reach-relevant environments. This is reminiscent of the “design stance” which human adopt toward artifacts, in which objects are understood in terms of what they were designed to do (Kelemen & Carey, 2007). According to this theory of conceptual formation, the underlying nature of an artifact is related to its intended function, which will constrain its form and materials, and provide the best explanatory variable for its appearance and construction. The present results suggest that this stance can explain reasoning about environments as well as artifacts. Indeed, function has also been found to act as an organizing principle for navigable-scale scenes,(Greene et al., 2016; Groen et al., 2018) suggesting that this is a general feature of our conceptual representation of things and spaces in the world. Altogether these results points to the possibility that the function of a space places strong constraints on its content and appearance, and to the general importance of goals for organizing our understanding of the world.

These results arise from human judgments obtained with relatively broad instructions. One open question is whether this global organization by function would appear even if participants received instructions to base their judgment on more specific, non-action factors. For example, would an embedding based specifically on judgments of location similarity, material property similarity, or visual similarity still retrieve function as a major organizing principle? It is possible that function organize high-level intuitive judgments, but other factors emerge for judgments based on lower-level factors, however, future work will be required to test this possibility.

5. Conclusion

Altogether, these findings point to distinct classes within the domain of reach-relevant environments. It is still an open question whether these distinctions are reflected in other aspects of reachspace perception. Future work is needed to establish whether these distinctions show additions dissociations in behavior, in their emergence across development, in their susceptibility to disruption following neurological events or cognitive decline, or in the large-scale neural activity they elicit. These results also have implications for the continuing study of reachspaces: as we develop theories of how reach-relevant environments are perceived and represented, it will be important to consider these distinctions and account for how they shape representations.

CRediT authorship contribution statement

Emilie L. Josephs: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Visualization, Writing – original draft, Writing – review & editing. **Martin N. Hebart:** Formal analysis, Methodology, Resources, Software, Validation, Visualization, Writing – review & editing. **Talia Konkle:** Conceptualization, Funding acquisition, Investigation, Project administration, Supervision, Validation, Visualization,

Writing – original draft, Writing – review & editing.

Data availability

Data will be placed in an OSF repository prior to publication

Acknowledgements

Work reported in this publication was supported by the National Eye Institute of the National Institutes of Health under Award Number R21EY031867, by a Pershing Square Fund grant for Research on the Foundations of Human Behavior, and by the Harvard Data Science Initiative. Thanks to the Konklab writing club for their feedback on early drafts of this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cognition.2023.105368.

References

Bartolo, A., et al. (2014). Contribution of the motor system to the perception of reachable space: An fmri study. *The European Journal of Neuroscience, 40*, 3807–3817.

Caramazza, A., & Shelton, J. R. (1998). Domain-specific knowledge systems in the brain: The animate-inanimate distinction. *Journal of Cognitive Neuroscience, 10*, 64.

Carroll, L. (2000). *Alice's adventures in wonderland.* Ontario: Broadview Press, 1832-1898.

Das, B., & Sengupta, A. K. (1996). Industrial workstation design: A systematic ergonomics approach. *Applied Ergonomics, 27*, 157–163.

Dobs, K., Isik, L., Pantazis, D., & Kanwisher, N. (2019). How face perception unfolds over time. *Nature Communications, 10*, 1–10.

Edelman, S. (1998). Representation is representation of similarities. *The Behavioral and Brain Sciences, 21*, 449–467. https://doi.org/10.1017/S0140525X98001253

Goldstone, R. L. (1994). The role of similarity in categorization: Providing a groundwork. *Cognition, 52*, 125–157.

Greene, M. R., Baldassano, C., Esteva, A., Beck, D. M., & Fei-Fei, L. (2016). Visual scenes are categorized by function. *Journal of Experimental Psychology. General, 145*, 82–94. https://doi.org/10.1037/xge0000129

Greene, M. R., & Oliva, A. (2006). Natural scene categorization from conjunctions of ecological global properties. *Proceedings of the Annual Meeting of the Cognitive Science Soceity, 28*, 7.

Greene, M. R., & Oliva, A. (2009). Recognition of natural scenes from global properties: Seeing the forest without representing the trees. *Cognitive Psychology, 58*, 137–176.

Groen, I. I., Silson, E. H., & Baker, C. I. (2017). Contributions of low-and high-level properties to neural processing of visual scenes in the human brain. *Philosophical Transactions of the Royal Society B, 372*, 20160102.

Groen, I. I., et al. (2018). Distinct contributions of functional and deep neural network features to representational similarity of scenes in human brain and behavior. *eLife, 7*. https://doi.org/10.7554/eLife.32962. e32962.

Hebart, M. N., Zheng, C. Y., Pereira, F., & Baker, C. I. (2020). Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. *Nature Human Behaviour, 4*, 1173–1185. https://doi.org/10.1038/s41562-020-00951-3

Hoffer, E., & Ailon, N. (2015). Deep metric learning using triplet network. In *International workshop on similarity-based pattern recognition* (pp. 84–92). Springer.

Hoyer, P. O. (2002). Non-negative sparse coding. In *Proceedings of the 12th IEEE workshop on neural networks for signal processing* (pp. 557–565). IEEE.

Hoyer, P. O. (2004). Non-negative matrix factorization with sparseness constraints. *Journal of Machine Learning Research, 5*.

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

Josephs, E. L., & Konkle, T. (2019). Perceptual dissociations among views of objects, scenes, and reachable spaces. *Journal of Experimental Psychology. Human Perception and Performance, 45*, 715–728. https://doi.org/10.1037/xhp0000626

Josephs, E. L., & Konkle, T. (2020). Large-scale dissociations between views of objects, scenes, and reachable-scale environments in visual cortex. *Proceedings of the National Academy of Sciences, 117*, 29354–29362. https://doi.org/10.1073/pnas.1912333117

Josephs, E. L., Zhao, H., & Konkle, T. (2021). The world within reach: An image database of reach-relevant environments. *Journal of Vision, 21*.

Jozwik, K. M., Kriegeskorte, N., Storrs, K. R., & Mur, M. (2017). Deep convolutional neural networks outperform feature-based but not categorical models in explaining object similarity judgments. *Frontiers in Psychology, 8*, 1726. https://doi.org/10.3389/fpsyg.2017.01726

Kelemen, D., & Carey, S. (2007). The essence of artifacts: Developing the design stance. *Creations of the mind: Theories of artifacts and their representation, 212–230*.

King, M. L., Groen, I. I., Steel, A., Kravitz, D. J., & Baker, C. I. (2019). Similarity judgments and cortical visual responses reflect different properties of object and scene categories in naturalistic images. *NeuroImage, 197*, 368–382.

Kirsh, D. (1995). The intelligent use of space. *Artificial Intelligence, 73*, 31–68.

Konkle, T., & Oliva, A. (2012). Canonical visual size for real-world objects. *Neuron, 37*, 1114–1124. https://doi.org/10.1037/a0020413

Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience, 2*, 4.

Landau, B., Smith, L. B., & Jones, S. S. (1988). The importance of shape in early lexical learning. *Cognitive Development, 3*, 299–321. https://doi.org/10.1016/0885-2014(88)90014-7

McRae, K., Cree, G. S., Seidenberg, M. S., & McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods, 37*, 547–559.

Mervis, C. B., & Rosch, E. (1981). Categorization of natural objects. *Annual Review of Psychology, 32*, 89–115. https://doi.org/10.1146/annurev.ps.32. 020181.000513

Morris, M. R., Brush, A. B., & Meyers, B. R. (2007). Reading revisited: Evaluating the usability of digital display surfaces for active reading tasks. In *Second annual IEEE international workshop on horizontal interactive human-computer systems (TABLETOP'07)* (pp. 79–86). IEEE.

Murphy, G. (2004). *The big book of concepts*. MIT press.

Oliva, A., & Torralba, A. (2006). Building the gist of a scene: The role of global image features in recognition. *Progress in Brain Research, 155*, 23–36.

Patterson, G., Xu, C., Su, H., & Hays, J. (2014). The sun attribute database: Beyond categories for deeper scene understanding. *International Journal of Computer Vision, 108*, 59–81. https://doi.org/10.1007/s11263-013-0695-z

Potvin, B., Swindells, C., Tory, M., & Storey, M.-A. (2012). Comparing horizontal and vertical surfaces for a collaborative design task. *Advances in Human-Computer Interaction, 2012*.

Previc, F. H. (1998). The neuropsychology of 3-d space. *Psychological Bulletin, 124*, 123.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology, 8*, 382–439.

Scott, S. D., Carpendale, M. S. T., & Inkpen, K. (2004). Territoriality in collaborative tabletop workspaces. In *Proceedings of the 2004 ACM conference on computer supported cooperative work* (pp. 294–303).

Shepard, R. (1987). Toward a universal law of generalization for psychological science. *Science, 237*, 1317–1323. https://doi.org/10.1126/science.3629243

Shepard, R. N., & Arabie, P. (1979). Additive clustering: Representation of similarities as combinations of discrete overlapping properties. *Psychological Review, 86*, 87.

Torralba, A., & Oliva, A. (2002). Depth estimation from image structure. In *, 24. IEEE Transactions on pattern analysis machine intelligence* (pp. 1226–1238).

Tversky, B. (1989). Parts, partonomies, and taxonomies. *Developmental Psychology, 25*, 983–995. https://doi.org/10.1037/0012-1649.25.6.983