scientific reports

Check for updates

OPEN

A data-driven investigation of human action representations

Diana C. Dima[1,2]✉, Martin N. Hebart[3] & Leyla Isik[1]

Understanding actions performed by others requires us to integrate different types of information about people, scenes, objects, and their interactions. What organizing dimensions does the mind use to make sense of this complex action space? To address this question, we collected intuitive similarity judgments across two large-scale sets of naturalistic videos depicting everyday actions. We used cross-validated sparse non-negative matrix factorization to identify the structure underlying action similarity judgments. A low-dimensional representation, consisting of nine to ten dimensions, was sufficient to accurately reconstruct human similarity judgments. The dimensions were robust to stimulus set perturbations and reproducible in a separate odd-one-out experiment. Human labels mapped these dimensions onto semantic axes relating to food, work, and home life; social axes relating to people and emotions; and one visual axis related to scene setting. While highly interpretable, these dimensions did not share a clear one-to-one correspondence with prior hypotheses of action-relevant dimensions. Together, our results reveal a low-dimensional set of robust and interpretable dimensions that organize intuitive action similarity judgments and highlight the importance of data-driven investigations of behavioral representations.

Our ability to rapidly recognize and respond to others' actions is remarkable, given the wide variety of human behaviors that span different contexts, goals, and motor sequences. When we see a person acting in the world, we integrate visual information, social cues and prior knowledge to interpret their action. These daily actions in context are often described as activities, which differ from other more basic-level or kinematic-based definitions of action, and despite their ubiquity, still pose a challenge to even state-of-the-art machine learning algorithms. How does the mind make sense of this complex action space?

Previous work on action understanding in the mind and brain has focused on hypothesis-driven efforts to identify critical action features and their neural underpinnings. This work has highlighted semantic content[1,2], social and affective features[3–5], and visual features[3,6] as essential components in visual action understanding. However, such an approach requires the experimenter to pre-define actions and their potential organizing dimensions, necessarily limiting the hypothesis space. Action categories have commonly been defined based on the verbs they represent[7] or everyday action categories as listed, for example, in the American Time Use Survey (ATUS)[3,5,8,9]. Given the diversity of actions, a low-dimensional, flexible representation may be a more efficient way to organize them in the mind and brain; but generating the hypotheses that could uncover this representation remains difficult, especially for naturalistic stimuli that vary along multiple axes.

Data-driven methods provide an alternative to pre-defined representational spaces and have achieved great success in mapping perceptual and psychological representations in other visual domains. In object recognition, a data-driven computational model revealed 49 interpretable dimensions capable of accurately predicting human similarity judgments[10]. Recent work has extended this method to near scenes, known as reachspaces, and identified 30 dimensions capturing their most important characteristics[11]. Low-dimensional representations have been also proposed that explain how people perceive others and their mental states[12,13] or psychologically meaningful situations[14,15].

To date there has been only limited data-driven work in the action domain. Using principal component analysis (PCA) of large-scale text data, a low-dimensional taxonomy of actions has been shown to explain neural data and human action judgments[16], as well as guide predictions about actions[17]. However, since this taxonomy was generated from text data, most of these dimensions were relatively abstract (e.g. *creation*, *tradition*, *spiritualism*), and it is unclear whether a similar set of dimensions would emerge from visual action representations. In the visual domain, six broad semantic clusters were shown to explain semantic similarity judgments of controlled action images[1], suggesting that actions may be semantically categorized at the superordinate level. However, it remains unclear how this finding would generalize to more natural and diverse stimulus sets.

[1]Department of Cognitive Science, Johns Hopkins University, Baltimore, USA. [2]Department of Computer Science, Western University, London, Canada. [3]Vision and Computational Cognition Group, Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany. ✉email: ddima@uwo.ca

We analyzed a dataset containing unconstrained behavioral similarity judgments of two sets of natural action videos from the Moments in Time dataset[18] collected in our prior study[5]. Behavioral similarity has often been used as a proxy for mental representations[19–21] and has been shown to correlate with neural representations[22–26]. Specifically, the perceived similarity of actions has been found to map onto critical action features, such as their goals or their social-affective content, as well as onto the structure of neural patterns elicited by actions[1,5,9].

Here, we employ a data-driven approach, sparse non-negative matrix factorization[27] (NMF) to recover the dimensions underlying behavioral similarity. This approach has two main advantages. First, it allows dimensions to be sparse, so that they need not be present in every action. For example, a single-agent action would have a value of 0 along a social interaction dimension. Second, the method requires the dimensions to be non-negative. Thus, dimensions can add up without canceling each other out, and no dimension can negate another's importance. Together, these criteria help recover interpretable dimensions, with values that are interpretable as the degree to which they are present in the data.

We show that a cross-validated approach to dimensionality reduction produces a low-dimensional representation that is interpretable by humans and generalizes across stimulus categories. Importantly, the dimensions recovered by NMF are more robust than those generated by the more commonly used PCA. The non-negativity constraint is known to yield a parts-based description, supporting dimension interpretability[28].

Using human labeling and semantic embeddings, we find that dimensions map to interpretable visual, semantic, and social axes and generalize across two experiments with different experimental structure, stimuli, and participants. Together, our results highlight the semantic structure underlying intuitive action similarity and show that cross-validated NMF is a useful tool for recovering interpretable, low-dimensional cognitive representations.

Results

NMF recovers robust dimensions. We analyzed two datasets consisting of three-second naturalistic videos of everyday actions from the Moments in Time dataset[18]. In two previously conducted experiments[5], participants arranged two sets of 152 and 65 videos from 18 everyday action categories[8] according to their unconstrained similarity[29]. The first dataset also included videos of natural scenes as a control category (see Stimuli; Supplementary Fig. 1; Supplementary Table 1).

During the experiments, participants arranged a maximum of 7–8 videos at a time inside a circular arena, and the task continued until sufficient evidence was obtained for each pair of videos[30] or until the experiment timed out (Experiment 1: 90 min; Experiment 2: 120 min). In Experiment 1, participants arranged different subsets of 30 videos from the 152-video set. In Experiment 2, participants arranged all 65 videos.

In both experiments, participants were instructed to arrange the videos according to how similar they were, thus allowing participants to use their own criteria to arrange the videos, as well as to use different criteria for different groupings of videos. This method allowed us to recover a multidimensional, intuitive representation of naturalistic actions.

We used sparse non-negative matrix factorization[27,31] with a nested cross-validation approach (see Methods) to recover the optimal number of underlying dimensions in the behavioral data (Fig. 1). This approach combines sparsity and non-negativity constraints to generate feature embeddings that can capture both categorical and continuous information[10,32,33] (see Methods). Using only behavioral similarity matrices as its starting point, this method can thus recover interpretable features that may shed light on how actions are organized in the mind.

Despite differences in stimulus set size and sampling, both experiments were characterized by similar numbers of dimensions (9 and 10 respectively; Supplementary Fig. 2) with a sparsity of 0.1. This suggests that the dimensions tended to be continuous and not categorical. Importantly, our sparse NMF procedure allowed the optimal structure to emerge from the data.

In Experiment 1, the final NMF reconstruction of the entire training set correlated well with the training data (Kendall's $\tau_A = 0.46$) and the held-out data ($\tau_A = 0.19$, true τ_A between the original training set and the hold-out

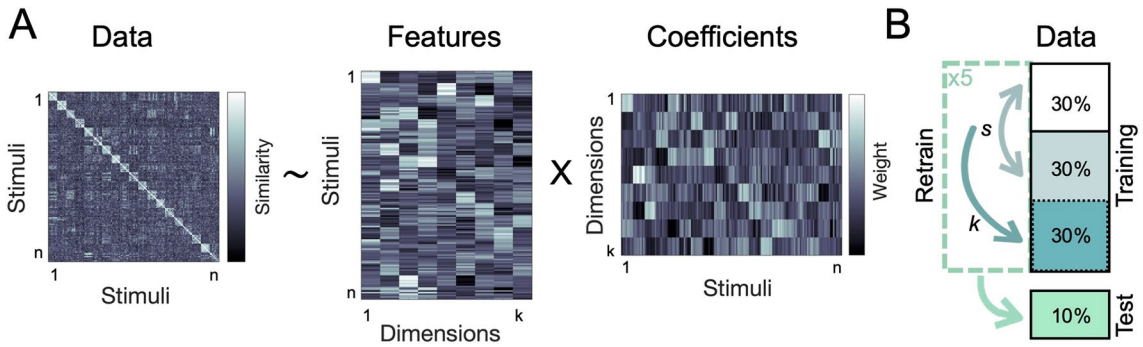

Figure 1. Analysis overview. (**A**) Using non-negative matrix factorization, we identified the optimal lower-dimensional approximation of a behavioral similarity matrix. This uncovered the interpretable dimensions underlying the perceived similarity of naturalistic action videos. (**B**) NMF cross-validation procedure. Individual similarity ratings were assigned to a cross-validation fold before averaging the input matrices for each fold. The sparsity parameters (s) were optimized using two-fold cross-validation on ~60% of the data, with a separate ~30% used to determine the number of dimensions (k), and a hold-out set of ~10% used for final evaluation.

set = 0.14). Performance was better in Experiment 2, with a training $\tau_A = 0.75$ and a hold-out $\tau_A = 0.46$ (true $\tau_A = 0.45$). In both experiments, the hold-out performance of NMF was close to the limit placed on it by the reliability of each dataset, as reflected in the true correlation between the training and hold-out sets.

Importantly, the dimensions were robust to systematic perturbations in the underlying stimulus sets (Fig. 2). Even after removing critical stimulus categories (such as all outdoor or indoor videos or certain action categories), the NMF procedure resulted in similar numbers of dimensions in both experiments (mean ± SD 8.4 ± 0.89 and 8.2 ± 1.64). All dimensions were significantly correlated to those resulting from the full stimulus set, suggesting that the NMF results generalize even after modifying the compositon of the underlying datasets.

NMF dimensionality varied less as a function of stimulus set size (average k range 6–8.3) than as a function of number of action categories (average k range 3.6–10.2; Supplementary Fig. 4). Further, NMF dimensions did not map directly onto any single visual, social, or action feature identified in our previous work[5] (Supplementary Fig. 3), suggesting that this method is able to capture additional information not revealed by a hypothesis-driven approach.

Finally, NMF performance was better than that achieved by an equivalent cross-validated analysis using PCA, which recovered 8 dimensions in both experiments (Experiment 1: training $\tau_A = 0.41$, hold-out $\tau_A = 0.16$; Experiment 2: training $\tau_A = 0.63$, hold-out $\tau_A = 0.41$). In the robustness analysis, the number of dimensions generated by PCA after removing critical stimulus categories was less reliable than those obtained with NMF in Experiment 1 (Experiment 1: 7.8 ± 2.49 vs. 8.4 ± 0.98; Experiment 2: 6 ± 1.58 vs 8.2 ± 1.64). While on average correlations with the original dimensions were high, their variance was also more than twice as high as that obtained with NMF (Supplementary Figs. 5–6). This suggests that dimensions recovered with PCA are more sensitive to variations in the underlying stimulus set than those found with NMF.

NMF recovers interpretable dimensions. The hypothesis-neutral dimensions generated by NMF suggest a potential structure to the behavioral space of action understanding. However, further validation is needed to show whether (1) these dimensions are reproducible and (2) to what degree they are interpretable.

To test reproducibility, participants in an online experiment selected the odd video out of a group consisting of seven highly weighted videos and one low-weighted video along each dimension. In a separate online experiment to test interpretability, participants were asked to provide up to three labels for each dimension after viewing the eight highest and eight lowest weighted videos. Their labels were quantitatively evaluated using FastText[34], a 300-dimensional word embedding pretrained on 1 million English words.

All dimensions were reproducible in the odd-one-out experiments (Fig. 3A; all $P < 0.004$), though participants performed significantly better on average in Experiment 1 (mean accuracy 0.8 ± 0.13) than in Experiment 2 (mean accuracy 0.61 ± 0.13, $t(15.82) = 3.69$, $P = 0.002$).

Participants' labels were consistent for most dimensions (Fig. 3B). Agreement, as measured via word embeddings, was higher in Experiment 1 (mean proportion 0.5 ± 0.2) than in Experiment 2 (mean proportion 0.34 ± 0.17), though this difference was not significant ($t(15.78) = 1.84$, $P = 0.08$).

The most common labels (Fig. 4) captured different types of information, ranging from visual (*nature/outdoors*), to action-related (*eating, cleaning, working*), as well as social and affective (*children/people, talking, celebration/happiness, chaos*). Dimensions in Experiment 2 included more social information overall, with four dimensions labeled with social or affective terms (*talking, people, celebration, chaos*), compared to one in Experiment 1 (*children*). Although many dimensions reflected action categories included in the dataset (*eating, cleaning,*

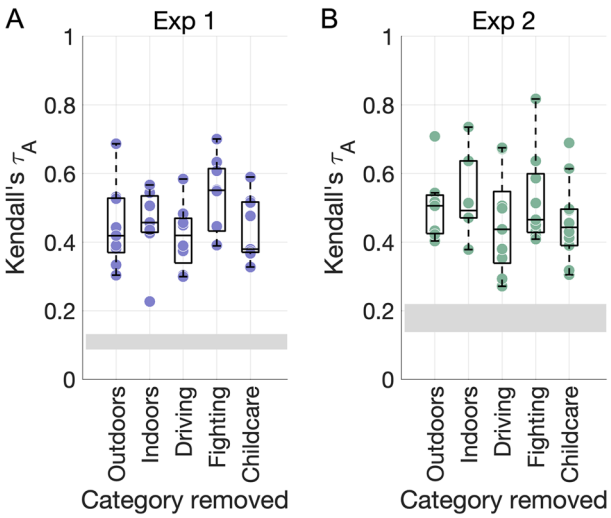

Figure 2. NMF dimension robustness. (**A**) The NMF procedure was repeated after removing key stimulus categories from the behavioral RDM from Experiment 1. Each dot shows the maximal correlation between each dimension obtained in the control analysis and any of the original dimensions with the same stimuli removed (repeats allowed). The grey rectangle depicts the chance level (min–max range). (**B**) As for (**A**), for Experiment 2.

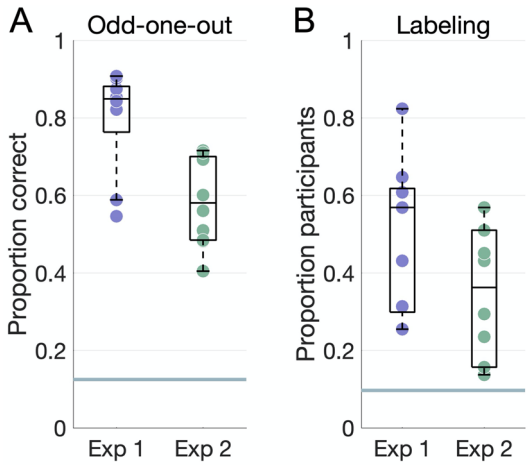

Figure 3. Behavioral results. (**A**) Accuracy on the odd-one-out task for each dimension plotted against the chance level of 12.5% (horizontal line). (**B**) Proportion of participants who agreed on the top label for each dimension, where agreement is defined as a word embedding dissimilarity in the 10th percentile within all dimensions in both experiments. The horizontal line marks a chance level based on embedding dissimilarity across different dimensions.

working, driving, reading) or labeled features that explained the most variance in our previous experiment (relating to people and affect), the information they provided was richer than the a priori category labels and crossed predefined category boundaries. For example, some videos were highly rated along several different dimensions (e.g. *work* and *learning*), thus capturing the complexity of naturalistic stimuli which often depict several actions or lend themselves to different interpretations.

Further, not all action categories were reflected in NMF dimensions, suggesting that certain action categories are more important than others in organizing behavior. Certain action categories were absorbed by others (e.g. *eating* included both *eating* and *preparing food*), while other related actions remained separated (e.g. *work* was split into *office work* vs *chores/cleaning*).

A shared semantic space. To better understand the relationship between dimensions revealed by the two datasets, we calculated Euclidean distances between averaged word embeddings for dimensions in each experiment (see Methods). This analysis revealed several dimensions that were present in both datasets: *eating, nature/outdoors, learning/reading, chores/cleaning,* and *work* (Fig. 5). Furthermore, some dimensions were moderately related to several others: *games: people, celebration; work: talking, working; reading: working, learning*. In Experiment 1, the only dimension that did not have a counterpart in Experiment 2 was *driving*, possibly because of the low number of driving videos in Experiment 2.

Discussion

Here, we used sparse non-negative matrix factorization to recover a low-dimensional representation of intuitive action similarity judgments across two naturalistic video datasets. This resulted in robust and interpretable dimensions that generalized across experiments. Our results highlight the visual, semantic and social axes that organize intuitive visual action understanding.

Non-negative matrix factorization as a viable approach to understanding similarity judgments. In the visual domain, it is reasonable to assume that features can be either absent or present to variable degrees, and that they can be additively combined to characterize a stimulus. Previous work has demonstrated that sparsity and positivity constraints enable the detection of interpretable dimensions underlying object similarity judgments[10]. Here, we showed that a different approach with the same constraints can recover robust, generalizable and interpretable dimensions of human actions. As opposed to those recovered for objects, the action dimensions were only moderately sparse, potentially due to the naturalistic nature of our stimuli. However, optimizing sparsity enabled us to strike the right balance between categorical and continuous descriptions of our data, thus capturing a rich underlying feature space[10,32,33].

Our approach recovered a similar number of dimensions across the two experiments (ten and nine), despite their different stimulus set sizes (152 vs. 65 videos). While the dimensions all had an interpretable, semantic description, none mapped directly onto previously used visual, semantic, or social features, suggesting that a data-driven approach can uncover additional information beyond hypothesis-driven analyses. Furthermore, the dimensions generalized across important stimulus categories like action category and scene setting (Fig. 2).

While a cross-validated PCA analysis uncovered a similar number of dimensions (eight), there was higher variance in the number and content of dimensions obtained after manipulating stimulus set composition (Supplementary Figs. 5–6). Visual inspection of the dimensions also suggested that they may be less interpretable than those uncovered by sparse NMF. For example, two dimensions in Experiment 1 appeared to depict driving

Figure 4. Label correspondence across experiments. Wordclouds showing the labels assigned by participants to each NMF dimension in Experiment 1 (left) and Experiment 2 (right), with larger font sizes representing more frequent labels. Bars connect dimensions from Experiment 1 to their most related dimensions from Experiment 2. The values shown are normalized relative similarities. Dimensions from Experiment 1 are sorted in descending order of their summed weights, while those from Experiment 2 are organized for clarity of visualization.

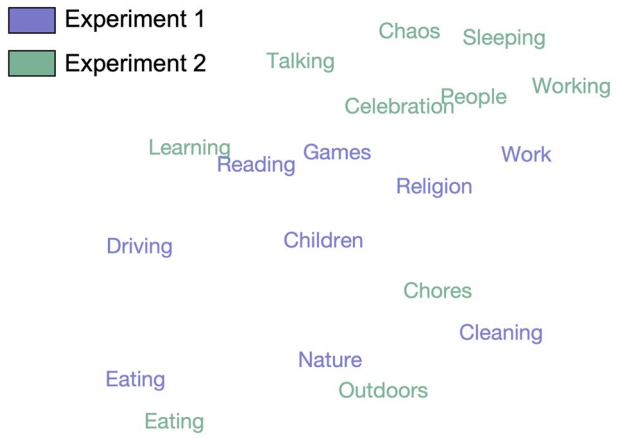

Figure 5. T-SNE plot displaying the distances between the averaged embeddings corresponding to each dimension from both experiments in a 2D space. *Eating*, *nature*, *cleaning*, *reading*, and *work* are the dimensions that most clearly replicate across experiments.

videos as the highest-weighted, yet these were interspersed with videos from different categories (e.g. cooking or socializing) that would make these dimensions difficult to label. The NMF *driving* dimension, on the other hand, showed the highest weights for the eight driving videos present in the dataset. Together, these results suggest that the positivity and sparsity constraints applied by NMF enable it to recover more robust and interpretable components from human behavioral data than PCA. These benefits are likely to extend to neural data, as suggested by the recent application of NMF to reveal novel category selectivity in human fMRI data[35].

Low-dimensional action representations. How should action categories be defined? This is a challenging question, particularly given neuroimaging evidence that actions are processed in the brain at different levels of abstraction[5,36,37]. Our results suggest that coarse semantic, visual, and social distinctions organize internal representations. Although we started with 18 activity categories, already defined at an arguably broad level, we find that our behavioral data is well-characterized by a lower number of broad dimensions.

The low dimensionality of the NMF reconstruction may seem surprising. Actions bridge visual domains, including scenes, objects, bodies and faces, and thus vary along a wide range of features. Furthermore, our use of naturalistic videos adds a layer of complexity compared to previous work using still images. However, a low-dimensional internal representation is more likely to enable the efficient and flexible action recognition that guides human behavior.

Mapping internal representations. We validated the resulting NMF dimensions in separate behavioral experiments. All dimensions were reproducible in an odd-one-out task (Fig. 3A) and consistently labeled by participants, as quantified through semantic embeddings (Fig. 3B). We visualized the most commonly assigned labels and assessed how they related to each other across the two experiments.

These analyses revealed several interpretable and reproducible dimensions, including those related to common everyday actions (*work*, *cleaning/chores*, *eating*, *reading/learning*), environment (*nature/outdoors*), and social information (*children/family, talking, people*). A previous data-driven analysis of semantic action similarity judgments found six clusters of actions related to locomotion, cleaning, food, leisure, and socializing[1]. Here, we found that some semantic categories emerged even in the absence of an explicit semantic task, while other dimensions reflected visual or social-affective features, highlighting the rich and varied information extracted from naturalistic actions.

Importantly, the NMF procedure did not simply return the action categories used to curate the dataset, and in fact none of the dimensions provided a one-to-one correspondence with semantic action category (Figs. 4 and 5). Instead, the dimension labels suggest that certain action categories were more salient than others (e.g. *work* or *eating*), while others tended to be grouped together based on other critical features, like scene setting or social structure.

For example, activities that take place outdoors, like hiking and certain sports, were grouped together under a *nature/outdoors* dimension. In Experiment 1, this dimension included control videos depicting natural scenes, while in Experiment 2, this dimension emerged in the absence of such control videos, suggesting that the natural environment is a salient organizing feature in itself (Fig. 4). While such scene-related information may not seem strictly action-related, recent proposals have suggested that these features may be critical for action understanding[38]. Indeed, scenes are often interpreted in terms of their affordance for action[39], and our work lends further support to these proposals.

Several dimensions were given labels pertaining to people (*children/family, talking, people*), highlighting the social structure of the similarity data revealed by our previous hypothesis-driven work[5]. In Experiment 2, videos depicting different actions were grouped together based on social or affective features like communication (talking face-to-face or on the phone) or negative affect (the *chaos* dimension, present, among others, in videos of people crying or fighting). These results are in line with previous work suggesting that social features,

including others' intentions and emotions, are important in action perception[5,9], and provide further insight into the specific social information that is prioritized.

The dimension labels revealed differences as well as similarities between the two experiments. Notably, dimensions in Experiments 2 included more social-affective information (Fig. 4), despite the fact that the two stimulus sets included the same action categories and were well-matched along social and affective dimensions[5]. However, the stimulus set in Experiment 2 was smaller, and stimulus sampling was conducted differently across the two experiments, resulting in more reliable similarity judgements in Experiment 2 (see Methods: Multiple arrangement). Despite these differences, the majority of dimensions correlated across experiments, suggesting that the NMF reconstructions form a shared semantic space, emerging in spite of stimulus set and sampling differences across experiments.

Neural underpinnings. Though the behavioral representations measured here likely reflect a late stage in action processing, they can reveal insights into the underlying neural representations. Key distinctions between our dimensions, such as the separation of person-directed (e.g., *talking* and *playing games*) versus object-directed (e.g., *chores* and *driving*) actions, are consistent with prior neural findings[4,6,38]. Sociality has also been identified as a key feature in neural action representations[3,5], as has information about the spatial layout of the environment[3].

However, the behavioral dimensions extracted here are finer grained than these broad distinctions, suggesting that specific object-directed actions or social content may be processed separately in the brain. These results, and large-scale data-driven experiments more generally, are a fruitful means of hypotehsis generation for future neural studies.

From actions to event representations. Naturalistic actions involve interactions between people, objects, and places, and it is thus no surprise that the dimensions we uncover reflect the richness of this information. This renders actions, as defined here, the ideal stepping stone towards higher-level event understanding. Another action taxonomy derived from data-driven text analysis proposed six broad action distinctions[16]; however, our dimensions are more concrete and specific, likely reflecting our input of visually depicted everyday human actions. Two dimensions (*food* and *work*) emerged in both the text data and our two video datasets. This opens exciting avenues for research into visual and language-based action understanding and whether they share a conceptual taxonomy.

Relatedly, stimulus selection is the biggest factor in determining the structure of similarity judgments. Here, both stimulus sets represented 18 everyday action categories based on the American Time Use Survey, curated so as to minimize visual confounds. These action categories may be described as activities or visual events, comprising sets of related actions that occur in daily life. While the number of stimuli does not impact the dimensionality of the final NMF reconstruction, the number of action categories does (Supplementary Fig. 3), and thus an accurate map of internal action representations will depend on comprehensive sampling of the relevant action space. Our results highlight a number of critical dimensions that organize how we judge the most common everyday actions; however, future research should expand this with datasets that sample actions in different ways, taking into account cultural and group differences in how we spend our time.

Together, our results highlight the low-dimensional structure that supports human action representations, and open exciting avenues for future research. Our stimuli and the resulting dimensions bridge the boundary between actions and situations, suggesting that our data-driven approach can be extended beyond specific visual domains to investigate how conceptual representations emerge in the mind and brain.

Methods

Stimuli. We analyzed two video datasets[5], each consisting of three-second naturalistic videos of everyday actions from the Moments in Time dataset[18].

The videos were selected to represent the following 18 common action categories based on the American Time Use Survey[8]: childcare; driving; eating; fighting; gardening; grooming; hiking; housework; instructing; playing games; preparing food; reading; religious activities; sleeping; socializing; sports; telephoning; and working. The dataset used in Experiment 1 included 152 videos, with 8 videos per action category and 8 control videos depicting natural scenes or objects. The dataset used in Experiment 2 included 65 videos, with 3–4 videos per action category. For more details, see Dima et al.[5].

Participants. We analyzed data from two previously conducted multiple arrangement experiments[5]. Experiment 1 involved 374 participants recruited via Amazon Mechanical Turk (300 after exclusions, located in the United States, gender and age not collected). 58 participants recruited through the Department of Psychological and Brain Sciences Research Portal at Johns Hopkins University took part in Experiment 2 (53 after exclusions, 31 female, 20 male, 1 non-binary, 1 not reported, mean age 19.38 ± 1.09).

Two experiments were conducted to validate the dimensions resulting from Experiments 1 and 2. 54 participants validated the dimensions from Experiment 1 (51 after exclusions, 33 female, 13 male, 1 non-binary, 4 not reported, mean age 19.25 ± 1.18) and a different set of 54 participants validated the dimensions from Experiment 2 (51 after exclusions, 37 female, 11 male, 3 not reported, mean age 20.12 ± 1.78). All subjects were recruited through the Department of Psychological and Brain Sciences Research Portal at Johns Hopkins University.

All procedures for online data collection were approved by the Johns Hopkins University Institutional Review Board, and informed consent was obtained from all participants. All research was performed in accordance with the Declaration of Helsinki.

Multiple arrangement. To measure the intuitive similarity between videos depicting everyday action events, we implemented a multiple arrangement task using the Meadows platform (www.meadows-research.com). Participants arranged the videos inside a circular arena according to their similarity. In order to capture intuitive, natural behavior, we did not define or constrain similarity. An adaptive algorithm ensured that different pairs of videos were presented in different trials, until a sufficient signal-to-noise ratio was achieved for each distance estimate. Behavioral representational dissimilarity matrices (RDM) were then constructed using inverse multi-dimensional scaling[30]. See Dima et al. 2022[5] for more details on the experimental procedure.

In Experiment 1, different subsets of 30 videos from the 152-video set were shown to different participants. The resulting behavioral RDM contained 11,476 video pairs with an average of 11.37 ± 3.08 ratings per pair.

In Experiment 2, participants arranged all 65 videos. The resulting behavioral RDM contained 2080 video pairs with 53 ratings per pair.

Non-negative matrix factorization (NMF). We used a data-driven approach, sparse NMF[27,31], to investigate the dimensions underlying action representations. This method has two important advantages over other forms of matrix decomposition, such as principal component analysis (PCA).

In aiming to represent each action video through a combination of underlying features, some of these may be assumed to be categorical. Such features would be present in some of the videos, but not in others, such that participants would arrange videos from the same category close together, and those outside the category farther apart. Sparse NMF applies sparsity constraints, allowing us to detect such categorical features that may group specific actions together.

However, the degree to which a feature is present may also distinguish certain actions from others, especially for features that capture non-categorical information. By enforcing positivity, NMF recovers continuous features with interpretable numerical values, reflecting the degree to which each feature is present in each stimulus. These two constraints thus allow both categorical and continuous structure to emerge, an approach well-suited to capture how real-world stimuli are represented in the mind[32,33].

Given a data matrix V, NMF outputs a basis vector matrix W and a coefficient matrix H with specified levels of sparsity and with k dimensions, such that $V \approx WH$. Since NMF can output different results when initialized with random matrices, we used non-negative singular value decomposition for initialization[40].

We first converted the behavioral RDM to a similarity matrix as used in symmetric applications of NMF[41]. As this matrix was symmetric, the output matrices were highly correlated (Pearson's $r > 0.93$), leading in practice to a similar solution to that given by symmetric NMF, where $W = H^T$.

We used a nested cross-validation scheme for NMF (Fig. 1B). In Experiment 1, in which different videos were arranged by different participants, cross-validation was implemented by leaving out randomly selected similarity ratings for each pair of videos; in Experiment 2, in which all participants arranged all videos, cross-validation was implemented by leaving out randomly selected participants.

Each training and test matrix used in cross-validation was created by averaging across similarity ratings (Experiment 1) or participants (Experiment 2). Due to the random sampling in Experiment 1, there were different numbers of ratings per video pair. Any missing datapoints after averaging (Experiment 1) were imputed (no more than 0.2% of any given similarity matrix). This was done by replacing each missing similarity value S using the following formula: $S_{a,b} = \max(\min(S_{a,b}, S_{a,c,} \ldots S_{a,n}), \min(S_{b,c}, S_{b,d}, \ldots S_{b,n}))$[42].

To evaluate the final performance of the NMF procedure, ~ 10% of the data was held out. In Experiment 1, this consisted of one randomly selected similarity rating for each pair of videos. The final test set was thus a complete similarity matrix with a single rating per pair (amounting to 9.52% of the data). In Experiment 2, the final test set consisted of five randomly selected participants' data (amounting to 9.43% of the data).

For parameter selection, the training data (~ 90% of all data) was divided into three sets (Fig. 1B).

We searched for the best sparsity parameters for each k (number of dimensions), up to 150 in Experiment 1 and 65 in Experiment 2 (just below the maximum number of videos in each experiment). The two sparsity parameters for W and H were selected using two-fold cross-validation on two thirds of the training data. In a hold-out procedure, the best combination of sparsity parameters for each k was tested on the remaining third of the training data. To speed up computation, we only tested combinations of sparsity parameters (s) ranging between 0 (no sparsity) and 0.8 (80% sparsity) in steps of 0.1. We selected the combination with maximal accuracy across the average of both folds, defined as the Kendall's τ_A correlation between the reconstructed WH matrix and the test matrix.

To increase robustness, this cross-validation procedure for sparsity parameter selection was repeated five times with different training set splits. The average performance curve on the held-out training set was used to select the best number of dimensions (k). To avoid overfitting, we identified the elbow point in this performance curve, defined as the point maximally distant from a line linking the two ends of the curve.

The NMF procedure was then reinitialized with the output of the first cross-validation fold and rerun on the whole training set (90% of the data) with the selected combination of parameters. The held-out 10% of the data was used to evaluate performance by calculating the Kendall's τ_A between the reconstructed NMF-based similarity matrix and the held-out test matrix.

Control analyses relating NMF dimensions to stimulus categories. We performed a post-hoc control analysis to assess the robustness of NMF dimensions to perturbations in the stimulus set. The NMF procedure was repeated after leaving out key stimulus categories that correlated with identified NMF dimensions (outdoors, indoors, childcare, driving, and fighting). To ensure these stimulus categories did not drive results, the dimensions obtained from each control analysis were correlated to the original dimensions. The correla-

tions were then tested against chance using one-tailed randomization testing with 1000 iterations of component matrix shuffling.

To evaluate whether NMF dimensions captured any obvious stimulus features (e.g. scene setting, action category or sociality), we assessed the correlation between each NMF dimension and 12 visual, action-related, and social features[5] (Supplementary Fig. 2).

Control PCA analysis. To asssess whether NMF provides an advantage over the more commonly used PCA, we conducted a similar cross-validated analysis using PCA, and assessed the resulting reconstruction accuracy and robustness to stimulus set perturbations in both experiments. The cross-validation procedure was exactly the same, except that no search for sparsity parameters was conducted. Instead, only the number of dimensions (k) was selected using two-fold cross-validation on the training data (~90% of the data).

Dimension validation. We used two tasks in two separate online experiments (corresponding to Experiments 1 and 2) to assess the interpretability of NMF dimensions in separate participant cohorts. We presented the eight highest weighted and eight lowest weighted videos along each dimension obtained from NMF as stimuli to the subjects. The experiment was implemented in JavaScript.

First, participants were asked to select the odd video out of a group consisting of seven highly weighted videos and one low-weighted video (odd-one-out) for a given dimension. This was done 20 times for each dimension with random resampling (from the top and bottom eight) of the videos shown. Participants were excluded if they did not achieve above-chance performance (over 12.5%) on catch trials involving a natural scene video as the odd-one-out among videos containing people. Dimensions were considered reproducible if participants achieved above-chance accuracy in selecting the odd-one-out (sign permutation testing, 5000 iterations, omnibus-corrected for multiple comparisons).

After completing this task, participants were asked to provide up to three labels (words or short phrases) for each dimension based on a visual inspection of the eight highest and eight lowest weighted videos.

Semantic analyses. We visually inspected the labels provided by participants to correct spelling errors and identify cases where pairs of antonyms were used to label a dimension (e.g. *nature vs home*); in these cases, we only kept the first label. Next, we visualized the labels by creating word clouds of the most common labels using the MATLAB *wordcloud* function.

To quantify participant agreement on labels, we used FastText[34], a 300-dimensional word embedding pretrained on 1 million English words. Embeddings were generated for each of the words and phrases provided by participants. Euclidean distances were then calculated across all labels within each dimension. Labels were considered related if the distance between them was in the 10th percentile across dimensions and experiments (below a threshold of $d = 1.2$). To generate a chance level for participant agreement, we calculated the proportion of related labels across different dimensions.

Finally, we assessed whether the NMF dimension labels replicated across the two experiments. To generate a dissimilarity matrix, embeddings were averaged across labels within each dimension before calculating Euclidean distances between dimensions. This allowed us to visualize which dimensions were most semantically related across experiments.

Data availability

Data related to this project is available as an Open Science Framework repository at https://osf.io/dxba7/. Analysis code is available on GitHub at https://github.com/dianadima/mot_nmf.

Received: 11 October 2022; Accepted: 23 March 2023
Published online: 30 March 2023

References

1. Tucciarelli, R., Wurm, M., Baccolo, E. & Lingnau, A. The representational space of observed actions. *Elife* **8**, 1–24 (2019).
2. Lingnau, A. & Downing, P. E. The lateral occipitotemporal cortex in action. *Trends Cognit. Sci.* **19**, 268–277 (2015).
3. Tarhan, L. & Konkle, T. Sociality and interaction envelope organize visual action representations. *Nat. Commun.* **11**, 1–11 (2020).
4. Wurm, M. F., Caramazza, A. & Lingnau, A. Action categories in lateral occipitotemporal cortex are organized along sociality and transitivity. *J. Neurosci.* **37**, 562–575 (2017).
5. Dima, D. C., Tomita, T. M., Honey, C. J. & Isik, L. Social-affective features drive human representations of observed actions. *Elife* **11**, e75027 (2022).
6. Wurm, M. F. & Caramazza, A. Lateral occipitotemporal cortex encodes perceptual components of social actions rather than abstract representations of sociality. *Neuroimage* **202**, 116153 (2019).
7. Bedny, M. & Caramazza, A. Perception, action, and word meanings in the human brain: The case from action verbs. *Ann. N. Y. Acad. Sci.* **1224**, 81–95 (2011).
8. ATUS. *American Time Use Survey*. *United States Department of Labor. Bureau of Labor Statistics* (2019).
9. Tarhan, L., De Freitas, J. & Konkle, T. Behavioral and neural representations en route to intuitive action understanding. *Neuropsychologia* **163**, 108048 (2021).
10. Hebart, M. N., Zheng, C. Y., Pereira, F. & Baker, C. I. Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. *Nat. Hum. Behav.* **4**, 1173–1185 (2020).
11. Josephs, E. L., Hebart, M. N. & Konkle, T. Emergent dimensions underlying human perception of the reachable world. *PsyArXiv* (2021).
12. Thornton, M. A. & Tamir, D. I. People represent mental states in terms of rationality, social impact, and valence: Validating the 3d Mind Model. *Cortex* **125**, 44–59 (2020).
13. Gray, H. M., Gray, K. & Wegner, D. M. Dimensions of mind perception. *Science (80-)* **315**, 619 (2007).

14. Rauthmann, J. F. *et al.* The situational Eight DIAMONDS: A taxonomy of major dimensions of situation characteristics. *J. Pers. Soc. Psychol.* **107**, 677–718 (2014).
15. Parrigon, S., Woo, S. E., Tay, L., Wang, T. & Wang, T. CAPTION-ing the situation: A lexically-derived taxonomy of psychological situation characteristics. *J. Pers. Soc. Psychol.* **112**, 642 (2016).
16. Thornton, M. A. & Tamir, D. I. Six dimensions describe action understanding: The ACT-FASTaxonomy. *J. Pers. Soc. Psychol.* **122**, 577–605 (2021).
17. Thornton, M. A. & Tamir, D. I. People accurately predict the transition probabilities between actions. *Sci. Adv.* **7**, eabd4995 (2021).
18. Monfort, M. *et al.* Moments in time dataset: One million videos for event understanding. *IEEE Trans. Pattern Anal. Mach. Intell.* **42**, 502–508 (2019).
19. Edelman, S. Representation is representation of similarities. *Behav. Brain Sci.* **21**, 449–498 (1998).
20. Shepard, R. N. Towards a universal law of generalization for psychological science. *Science* **80**(237), 1317–1323 (1987).
21. Murphy, G. L. *The Big Book of Concepts* (MIT Press, Cambridge, 2002).
22. Charest, I. *et al.* Unique semantic space in the brain of each beholder predicts perceived similarity. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 14565–14570 (2014).
23. Cichy, R. M., Kriegeskorte, N., Jozwik, K. M., van den Bosch, J. J. F. & Charest, I. The spatiotemporal neural dynamics underlying perceived similarity for real-world objects. *Neuroimage* **194**, 12–24 (2019).
24. Wardle, S. G., Kriegeskorte, N., Grootswagers, T., Khaligh-Razavi, S. M. & Carlson, T. A. Perceptual similarity of visual patterns predicts dynamic neural activation patterns measured with MEG. *Neuroimage* **132**, 59–70 (2016).
25. Proklova, D., Kaiser, D. & Peelen, M. V. MEG sensor patterns reflect perceptual but not categorical similarity of animate and inanimate objects. *Neuroimage* **193**, 167–177 (2019).
26. Bankson, B. B., Hebart, M. N., Groen, I. I. A. & Baker, C. I. The temporal evolution of conceptual object representations revealed through models of behavior, semantics and deep neural networks. *Neuroimage* **178**, 172–182 (2018).
27. Hoyer, P. O. Non-negative matrix factorization with sparseness constraints. *J. Mach. Learn. Res.* **5**, 1457–1469 (2004).
28. Lee, D. D. & Seung, H. S. Learning the parts of objects by non-negative matrix factorization. *Nature* **401**, 788–791 (1999).
29. Goldstone, R. An efficient method for obtaining similarity data. *Behav. Res. Methods Instrum. Comput.* **26**, 381–386 (1994).
30. Kriegeskorte, N. & Mur, M. Inverse MDS: Inferring dissimilarity structure from multiple item arrangements. *Front. Psychol.* **3**, 1–13 (2012).
31. Hoyer, P. O. Non-negative sparse coding. *Neural Netw. Signal Process. Proc. EEE Work.* https://doi.org/10.1109/NNSP.2002.1030067 (2002).
32. 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).
33. Zheng, C. Y., Baker, C. I., Pereira, F. & Hebart, M. N. Revealing interpretable object representations from human behavior. In *7th Int. Conf. Learn. Represent. ICLR 2019* 1–16 (2019).
34. Bojanowski, P., Grave, E., Joulin, A. & Mikolov, T. Enriching word vectors with subword information. *Trans. Assoc. Comput. Linguist.* **5**, 135–146 (2016).
35. Khosla, M., Ratan Murty, N. A. & Kanwisher, N. A highly selective response to food in human visual cortex revealed by hypothesis-free voxel decomposition. *Curr. Biol.* https://doi.org/10.1016/j.cub.2022.08.009 (2002).
36. Iordan, M. C., Greene, M. R., Beck, D. M. & Fei-Fei, L. Basic level category structure emerges gradually across human ventral visual cortex. *J. Cognit. Neurosci.* **27**, 1427–1446 (2015).
37. Spunt, R. P., Kemmerer, D. & Adolphs, R. The neural basis of conceptualizing the same action at different levels of abstraction. *Soc. Cognit. Affect. Neurosci.* **11**, 1141–1151 (2016).
38. Wurm, M. F. & Caramazza, A. Two 'what' pathways for action and object recognition. *Trends Cognit. Sci.* **26**, 103–116 (2022).
39. Greene, M. R., Baldassano, C., Esteva, A., Beck, D. M. & Fei-fei, L. Visual scenes are categorized by function. *J. Exp. Psychol. Gen.* **145**, 82–94 (2016).
40. Boutsidis, C. & Gallopoulos, E. SVD based initialization: A head start for nonnegative matrix factorization. *Pattern Recognit.* **41**, 1350–1362 (2008).
41. Kuang, D., Ding, C. & Park, H. Symmetric nonnegative matrix factorization for graph clustering. In *Proc. 12th SIAM Int. Conf. Data Mining, SDM 2012* 106–117 (2012). https://doi.org/10.1137/1.9781611972825.10.
42. Lapointe, F. J. & Kirsch, J. A. W. Estimating phylogenies from lacunose distance matrices, with special reference to DNA hybridization data. *Mol. Biol. Evol.* **12**, 266–284 (1995).

Acknowledgements

The authors would like to thank Christopher Honey and Tyler Tomita for their contribution to the action dataset and their analysis suggestions.

Author contributions

D.C.D.: conceptualization, methodology, software, analysis, investigation, visualization, writing—original draft, writing—review and editing; M.N.H.: conceptualization, methodology, software, writing—review and editing; L.I.: conceptualization, funding acquisition, methodology, resources, writing—review and editing. All authors have reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-023-32192-5.

Correspondence and requests for materials should be addressed to D.C.D.