The ventral visual pathway represents animal appearance over animacy, unlike human behavior and deep neural networks

Stefania Bracci[1], Ioannis Kalfas[2], and Hans Op de Beeck[1]

(1) Laboratory of Biological Psychology, KU Leuven, Leuven, 3000, Belgium

(2) Laboratory of Neuro- and Psychophysiology, Department of Neurosciences, KU Leuven, Leuven, 3000, Belgium

Corresponding author:

stefania.bracci@kuleuven.be

Laboratory for Biological Psychology

Tiensestraat 102 - box 3714

3000 Leuven

Tel. +32 16 37 77 44

Running Title: DNNs predict human perception but not brain representations

Abstract

Recent studies showed general agreement between object representations in the ventral visual pathway and representational similarities in behavior and deep neural networks. In this fMRI study we challenge this state-of-the-art by dissociating object appearance (how does the object look like?) from object category (which object category is it?). The stimulus set includes animate objects (e.g., a cow), typical inanimate objects (e.g., a mug), and, crucially, inanimate objects that look like the animate objects (e.g., a cow-shaped mug). Behavioral judgments and deep neural networks showed a strong effect of the animacy dimension, setting the lookalike (and inanimate) objects apart from the animate ones. In contrast, neural activity patterns in ventral occipitotemporal regions were strongly biased towards object appearance: animate entities and lookalikes were similarly represented, and separated from the regular inanimate objects. Thus, despite its importance for human behavior and neural networks, animacy is not well represented in ventral visual cortex.

Keywords

fMRI, MVPA, deep neural networks, object representations, animacy, ventral occipitotemporal cortex.

Introduction

Over the years, studies in cognitive and systems neuroscience have characterized object representations in the primate brain in terms of their distinctive properties. Early behavioral studies already demonstrated the role of multiple visual features and visual similarity in categorization and recognition (e.g., Kruschke, 1992; Medin and Schaffer, 1978; Nosofsky, 1984) providing important information about the type of features and formats characterizing the underlying representations (Biederman, 1987; Tarr and Bulthoff, 1998). The neural mechanisms became accessible through the discovery of focal and distributed representations for superordinate distinctions for animate and inanimate objects, and for more specific distinctions such as faces, bodies, and other object categories (e.g., Haxby et al., 2001; Kanwisher et al., 1997; Martin et al., 1996). Since 2000, monkey neurophysiology and human imaging studies, unraveled the underlying neural representations in terms of their representational similarity properties (e.g., Kiani et al., 2007; Kriegeskorte and Kievit, 2013; Kriegeskorte et al., 2008b; Op de Beeck et al., 2001) as well as the role of a wide range of visual and semantic object properties such as shape, size, and animacy (e.g., Bracci and Op de Beeck, 2016; Konkle and Oliva, 2012; Nasr et al., 2014).

Recent computational studies have lifted our understanding of these multidimensional and hierarchical object representations to a higher level by providing a type of models that are able to reach human behavioral performance in the categorization of complex images (He et al., 2015; Kheradpisheh et al., 2016a; Krizhevsky et al 2012; Russakovsky et al., 2014). It has been shown that deep convolutional networks trained in object recognition tasks develop representations that share many of the properties of primate object representations. For example, there is accumulating evidence of object representational similarity correspondence among artificial networks, human judgments, and brain representations (e.g., Cadieu et al., 2014; Cichy, et al., 2016; Guclu and van Gerven, 2014; Guclu and van Gerven 2015; Kalfas et al., 2017; Khaligh-Razavi and Kriegeskorte, 2014; Kubilius et al., 2016; Yamins et al., 2014). This correspondence extends to the representation of several object dimensions, such as shape properties (Kubilius et al., 2016), and the distinction between animate and inanimate objects (Khaligh-Razavi et al., 2014). The availability of computational models that can mimic human recognition behavior and neural information processing in its full complexity offers exciting possibilities for understanding human object vision (Kriegeskorte, 2015).

In this human functional neuroimaging study we challenge this state of the art by providing a critical test of the superordinate categorical distinction of animate versus inanimate, which seems to be an important and maybe even the dominant dimension in ventral visual cortex and in neural networks (Kriegeskorte et al., 2008b). The aforementioned studies did not intentionally dissociate object identity from object appearance, two dimensions that are typically closely correlated in natural images. To overcome this, we constructed a stimulus set that was explicitly designed to dissociate object appearance (how an object looks like) from object category (what the object really is), with an emphasis on animacy. We show that human participants can very well say what the object is and whether it is animate or inanimate, and so can advanced deep neural networks (DNNs). However, the neural activity patterns in category-selective cortex were confused by animal appearance. For the human ventral visual cortex, a cow-shaped mug is more similar to a cow than to a mug. This effect was surprisingly strong and replicated in two different task contexts, persisting even when subjects were engaged in a task requiring them to explicitly report that the cow-shaped mug was an inanimate object. In contrast, animacy was represented to a much higher degree in human behavior and in deep neural networks. As a consequence, rather surprisingly, deep neural networks, which were never explicitly trained on animacy,

outperform ventral occipitotemporal cortex representations in categorizing objects as being animate or inanimate.

Results

We constructed a stimulus set to intentionally separate object appearance from object identity, including: 9 animals (e.g., a cow), 9 objects (e.g., a mug), and 9 lookalike objects, which consisted of stimuli that closely match the identity of one of the inanimate objects but looked like one of the animals (e.g., a cow-shaped mug). This resulted in a stimulus set of 9 closely matched triads -- 27 stimuli in total (Figure 1A). The mug and the cow-shaped mug are identical in all possible respects except their appearance (e.g., object identity, functional properties, size, material made of, manipulability knowledge, and animacy). Conversely, the cow-shaped mug and the cow differ in all the above semantic and functional properties but share animal appearance.

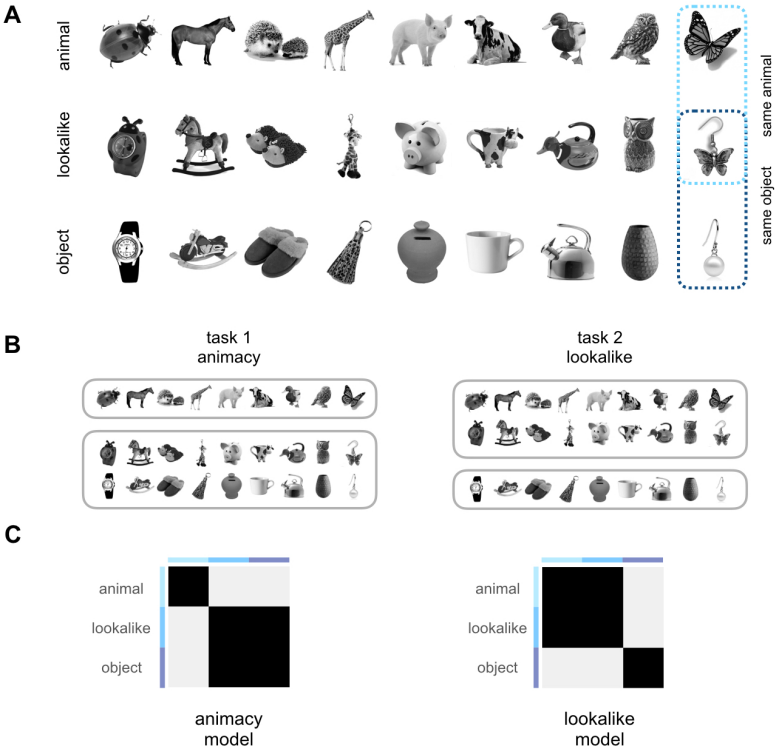


Figure 1. The experimental design. (A) The stimulus set was specifically designed to dissociate object appearance from object identity. We included 9 different object triads. Each triad included an animal (e.g., butterfly), an inanimate object (e.g., earring), and a lookalike object closely matched to the inanimate object in terms of object identity and to the living animal in terms of object appearance (e.g., a butterfly-shaped earring). (B) During fMRI acquisition, participants performed two different tasks counterbalanced across runs. During the *animacy task* (task 1) participants judged animacy ("does this image depict a living animal?"). During the *lookalike task* (task 2) participants judged animal appearance ("does this image look like an animal?"). Participants responded 'yes' or 'no' with the index and middle finger. Responses were counterbalanced across runs. (C) Model predictions based on the two tasks. The animacy model predicts high similarity between the object and the lookalike condition, thus predicting all inanimate objects (objects and lookalikes) to cluster separately from living animals. On the contrary, the lookalike model predicts high similarities between the animal and the lookalike condition despite differences in object identity, functional properties, size, and animacy. The two models are independent ($r = 0.07$).

Deep neural networks and human perception predict a representation based on animacy.

We compared human similarity judgments (without specifying which dimension to pay attention to; see Methods) to the stimulus representation in two DNNs (VGG-19 and GoogLeNet) that were previously shown to match human performance (He et al., 2015; Kheradpisheh et al., 2016b; Russakovsky et al., 2014). The object dissimilarity for each image pair (1 minus Pearson's r) was computed for human similarity judgments (Methods), and for the output vectors of the DNNs' upper layer (Methods), and resulting dissimilarity matrices were

correlated with the two independent models. Results revealed similarities but also differences between human judgments and DNNs representations (Figure 2). A significant positive correlation of both models (lookalike: $t_{(14)} = 4.11$, $p = 0.001$; animacy: $t_{(14)} = 3.63$, $p = 0.003$) with behavioral judgments showed that participants perceived both dimensions, with a slight bias for animacy; the correlation with the animacy model was numerically higher than the lookalike model ($t_{(14)} = 1.41$, $p = 0.18$). This bias was particularly strong in DNNs' representations, with a high correlation with the animacy model (VGG-19: $r = 0.73$; $p < 0.00001$; GoogLeNet: $r = 0.55$; $p < 0.00001$), but no significant correlation with the lookalike model (VGG-19: $r = -0.04$; GoogLeNet: $r = -0.01$; $p > 0.5$, for both DNNs). These results are also evident when inspecting dissimilarity matrices and 2D arrangements derived from multidimensional scaling (MDS; Figure 3A-B). Animals do not mingle with objects and lookalikes in any of these spatial configurations. In addition, objects and lookalikes are cluttered together in DNNs representations, while they appear to be partially separated in human similarity judgments. Taken together, these data show that both humans and DNNs set apart animate from inanimate objects in accordance with one of the most reported divisions in ventral occipito-temporal cortex (VTC) (Grill-Spector and Weiner, 2014; Kriegeskorte et al., 2008b).

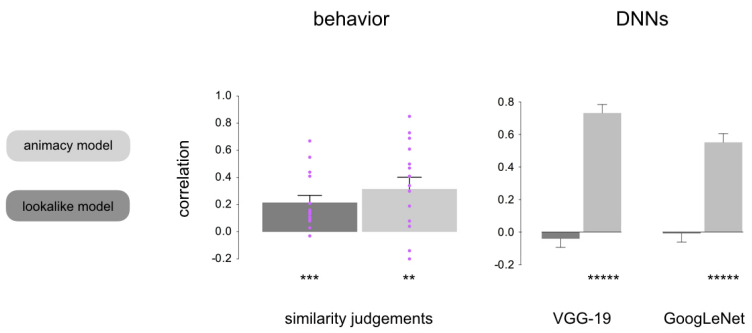


Figure 2. Similarities and differences between DNNs representations and human judgments. Representational similarity analysis (RSA; Kriegeskorte et al., 2008a) results for the lookalike model (dark gray) and the animacy model (light gray) are shown for human judgments (left panel), and DNNs (right panel). For behavioral results, individual participant's correlation values (n=15) are shown in purple. Asterisks indicate significant values relative to baseline, and error bars indicate SEM. For DNNs results, asterisks indicate significant values computed with permutation tests (10,000 randomizations of stimulus labels), and error bars indicate SE computed by bootstrap resampling of the stimuli. ***** $p < 0.00001$, **** $p < 0.0001$, *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

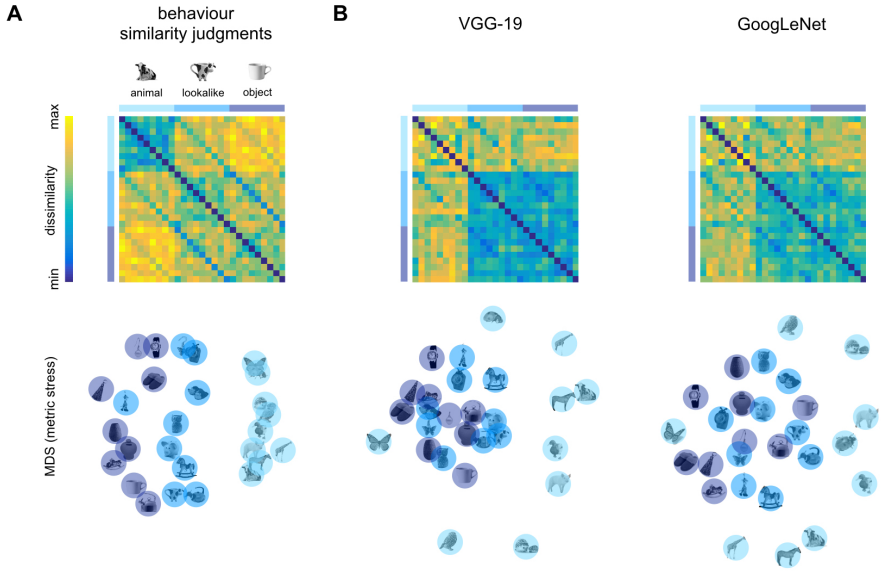


Figure 3. Deep neural networks and human perception predict a representation based on object animacy. (A-B) Dissimilarities matrices (upper panel) and 2D arrangements derived from MDS (metric stress; lower panel) show stimuli pairwise distances for behavioral judgments (A) and DNNs (B). Light blue: animals; blue: lookalikes; dark blue: objects.

Animal lookalike explains representational content in human visual cortex

The above results, from DNNs and human behavior, point to the animacy model as the representational structure to sustain object categorization behavior even with a stimulus set that sets apart object identity from object appearance. Based on previous studies, we would expect this organizational principle to have its neural correlate in VTC (Grill-Spector and Weiner, 2014; Kriegeskorte et al., 2008b), possibly accompanied by a further effect of visual appearance (Bracci and Op de Beeck, 2016). We collected human functional neuroimaging (fMRI) data in an event-related design (Methods). All 27 images were presented in a random order while participants performed two different tasks counterbalanced across runs (Figure 1B). During the animacy task (task 1), participants judged whether the image on the screen depicted a living animal (yes or no). During the lookalike task (task 2), participants judged whether the image on the screen looked like an animal (yes or no). In this way, we forced participants to group the lookalike condition in two different ways depending on their properties: either similarly to the object condition (as in the animacy model, see Figure 1C) or similarly to the animal condition (as in the lookalike model). Thus, in addition to test the two independent predictive models, we can assess any task-related modulation.

We correlated the dissimilarity values predicted by the two models with the dissimilarity matrices derived from neural activity patterns elicited in target regions of interest (ROIs; Figure 4A) in visual cortex. Our main ROI was VTC, divided into the anterior portion (ant VTC) and the posterior portion (post VTC). We also included V1 as a control ROI. Results were analyzed separately for each task but since no task-related effects were observed, we mostly report results for the full dataset in the main text. Statistics for the results analyzed for the two tasks separately are reported in the Supplementary Results.

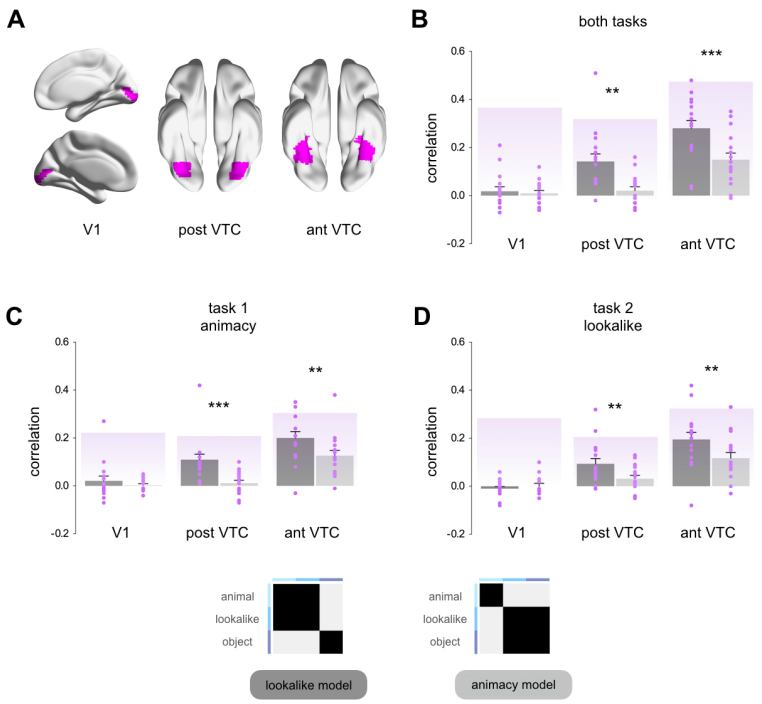


Figure 4. Animal lookalike explains representational content in human visual cortex. (A) Group averaged ROIs are shown on an inflated human brain template in BrainNet Viewer (Xia et al., 2013). (B-D) Representational similarity analysis (RSA; Kriegeskorte et al., 2008) results for the lookalike model (dark gray) and the animacy model (light gray) are shown for the data combined across the two tasks (B), and for each task separately (C-D). Individual participant's (n =16) correlation values are shown in purple. Purple-shaded backgrounds represent reliability values of the correlational patterns taking into account the noise in the data. These values give an estimate of the highest correlation we can expect in each ROI. Error bars indicate SEM. Asterisks indicate significant difference between the two models (*** $p < 0.001$; ** $p < 0.01$).

To investigate the role of animacy and animal appearance in driving the VTC organization, correlation values were tested in a 3x2 ANOVA with ROI (V1, post VTC, ant VTC) and Model (lookalike, animacy) as within-subject factors. Results revealed a significant ROI x Model interaction ($F_{(2,30)} = 9.40$, $p = 0.001$; Figure 4B), thus highlighting differences in the relation between the two models and representational content in the three ROIs. No positive correlations were found in V1 (lookalike: $t < 1$; animacy: $t < 1$), suggesting that our stimulus set was constructed appropriately to investigate neural representations without trivial confounds with low-level visual features. Unexpectedly, and differently from DNNs and behavioral results, neural representations in anterior and posterior VTC were significantly more correlated with the lookalike model than the animacy model (post-VTC: $t_{(15)} = 3.85$, $p = 0.002$; ant-VTC: $t_{(15)} = 4.00$, $p = 0.001$). In addition, we also observed differences between the anterior and posterior portion of VTC. Whereas in anterior VTC both models were significantly correlated with the neural data (lookalike: $t_{(15)} = 8.70$, $p < 0.00001$; animacy: $t_{(15)} = 5.36$, $p < 0.00001$), in posterior VTC, only correlations with the lookalike model significantly differed from baseline (lookalike: $t_{(15)} = 4.56$, $p = 0.0004$; animacy: $t < 1.5$). To visualize the representational structure in each ROI in more detail, dissimilarities matrices and 2-dimensional arrangements derived from MDS are shown in Figure 5A-B. Findings from the ROI-based analysis were backed up by a whole-brain searchlight analysis, which showed a significant preference for the lookalike model over the animacy model in (right) VTC (Supplementary Figure1; Supplementary Results).

Replicating results within subjects, data analyzed for the two tasks separately did not reveal any task modulation in VTC (Figure 3C-D; Supplementary Results). In both tasks, the lookalike model was significantly more correlated with the neural data than the animacy model in post VTC (task 1: $t_{(15)} = 3.83$, $p = 0.002$; task

2: $t_{(15)} = 3.05$, $p = 0.008$) and ant VTC (task 1: $t_{(15)} = 2.97$, $p = 0.009$; task 2: $t_{(15)} = 3.44$, $p = 0.004$). Furthermore, none of the ROIs showed an effect of task in a direct statistical comparison (VTC-ant: $F < 1$; VTC-post: $F < 1$).

Together these results show that the animacy organization reported in VTC might be largely explained by animal appearance rather than of object animacy. Inanimate (lookalike) objects that share with animals neither functional properties nor animacy, are represented closer to animals than to objects with which they share the same object category (e.g., a mug). The animal appearance might relate to visual features present in faces, eyes, limbs, and bodies, which are not generally present in inanimate objects. In agreement, those animals whose face (or body) looks less prototypical (e.g., the butterfly and the ladybug in Figure 5B) seemed to be represented closer to objects than animals (Sha et al., 2015). This result, replicated across tasks, is particularly surprising in light of results from human judgments and DNNs, which privilege object category over appearance.

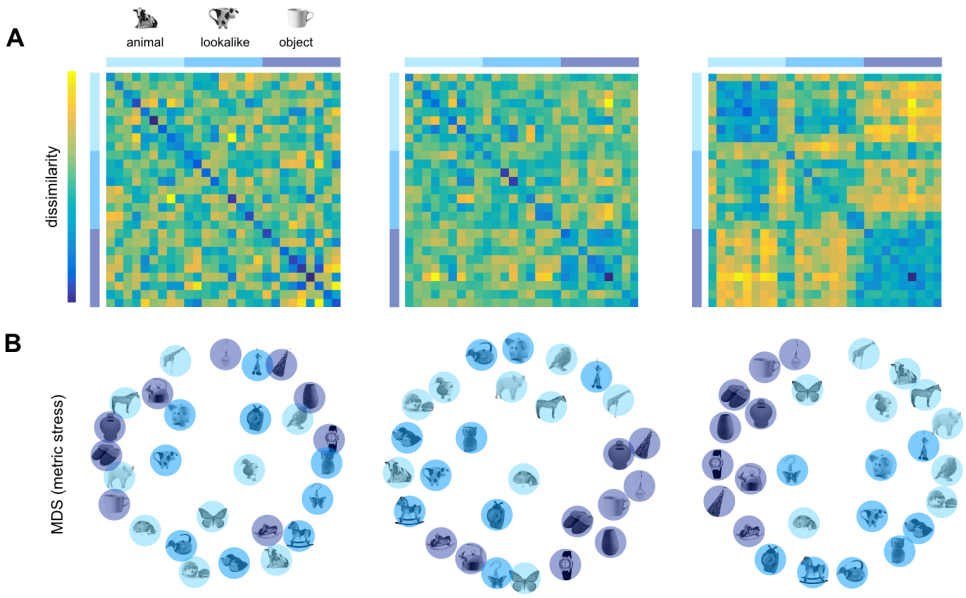


Figure 5. The neural similarity space in VTC reflects animal appearance. (A) The neural dissimilarity matrices (1 minus Person's r) derived from the neural data (averaged across subjects and tasks) show pairwise dissimilarities among stimuli in the three ROIs. (B) The MDS (metric stress), performed on the dissimilarity matrices averaged across subjects and tasks, shows pairwise distances in a 2D space for the three ROIs. Animals: light blue; lookalike: blue; and objects: dark blue.

Do VTC representations distinguish between lookalikes and real animals?

In anterior VTC the two (independent) models both differ from baseline. Therefore, although the lookalikes were closer to animals than to objects, lookalikes and animals might be encoded separately in VTC. Another possibility is that the positive correlation with the animacy model is fully driven by the distinction between animals and (nonlookalike) objects (which both predictive models share), without any representation of the lookalikes as being different from real animals (on which the two predictive models make opposite predictions). In other words, does VTC discriminate between animals and objects that look like animals or is it largely blind to this category distinction? If VTC does not distinguish between animals and objects that have animal appearance then there should be no difference between within- and between-condition correlations for these two categories. In what follows we investigated this hypothesis.

For each subject and condition, we computed the average within-condition correlation (e.g., comparing an animal with animals from other triads) and subtracted the averaged between-condition correlation (e.g., comparing the animal with the lookalikes or with the objects from other triads). We call this the *category index*. We computed the *category index* for each condition pair (i.e., animals-lookalikes, animals-objects, and lookalikes-objects). Category indexes above baseline indicate that two category representations are separated. In both VTC regions, the category index could distinguish between animals and objects (ant VTC: $t_{(15)} = 6.01$, $p = 0.00002$; post VTC: $t_{(15)} = 3.76$, $p = 0.002$; Figure 6, left panel), and objects and lookalikes (ant VTC: $t_{(15)} = 7.48$, $p < 0.00001$; post VTC: $t_{(15)} = 3.84$, $p = 0.002$). The category index between animals and lookalikes was significant in ant VTC ($t_{(15)} = 4.28$, $p = 0.001$) but did not differ from baseline in post VTC ($t < 1.5$). Furthermore, in both ROIs, the category index for animals and lookalikes was significantly smaller that the other two category indexes (ant VTC: $t_{(15)} = 4.00$, $p = 0.002$; post VTC: $t_{(15)} = 3.53$, $p = 0.004$). In V1 none of the conditions could be distinguished ($t < 1$, for all condition pairs). As for previous analyses, these results were replicated when data was analyzed for each task separately (Supplementary Figure 2). Thus, VTC representations reflect animal appearance much more than animacy, with a small remaining difference between animals and lookalikes in the anterior part of VTC (but not posterior VTC).

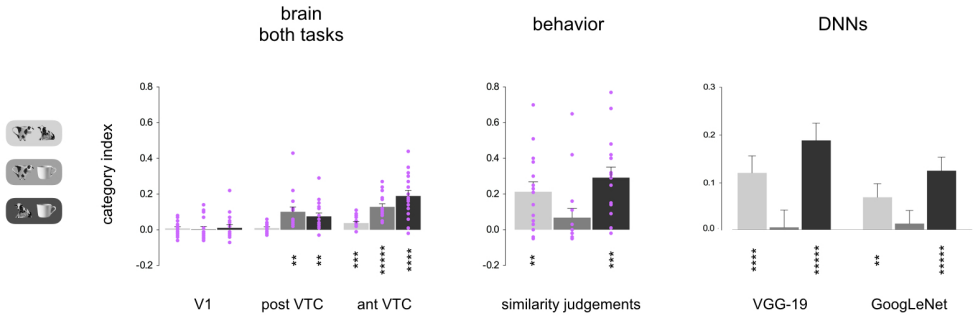


Figure 6. Differential category discriminability information in VTC and DNNs**.** The category index reflects representational discriminability among the three stimulus categories (animals, lookalikes, and objects) and is computed for each condition pair by subtracting the average of between-condition correlations (e.g., for animals and lookalikes) from the average of within-condition correlations (e.g., for animals and lookalikes). Results are reported for neural data (left panel), behavior (middle panel), and DNNs (right panel). Light gray: animals-lookalikes; gray: lookalikes-objects; dark gray: animals-objects. For neural and behavioral data, individual participant's (n=16) data points are shown in purple. Asterisks indicate significant values relative to baseline, and error bars indicate SEM. For DNNs data, asterisks indicate significant values computed with permutation tests (10,000 randomizations of stimulus labels), and error bars indicate SE computed by bootstrap resampling of the stimuli. ***** $p < 0.00001$, **** $< p < 0.0001$, *** $< p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

Opposite results were observed for behavioral judgments and DNNs. In both cases, the category index was highly significant for the distinction between animals and objects (behavior: $t_{(14)} = 4.90$, $p < 0.00001$; VGG-19: $p = 0.00001$; GoogLeNet: $p < 0.00001$), and between animals and lookalikes (behavior: $t_{(14)} = 3.77$, $p = 0.002$; VGG-19: $p = 0.00002$; GoogLeNet: $p = 0.008$), but did not differ between lookalikes and objects (behavior: $t < 1.5$; VGG-19: $p = 0.4$; GoogLeNet: $p = 0.30$). This suggests that differently from VTC representations, human judgments and convolutional neural network take object identity into account; two mugs belong to the same category regardless of their shape.

Together, these results show that representations in VTC are not predicted either by convolutional neural networks or human judgments. Whereas the former privileges the role of animal appearance, the latter privilege the role of basic object categories.

The coding for animal appearance interferes with appearance-invariant identification.

Previous studies have shown that VTC contains information about object identity, with a high degree of invariance for a variety of image transformations (Desimone et al., 1984; Grill-Spector et al., 1999; Grill-Spector et al., 1998; Gross et al., 1972). Despite the reported finding that a lookalike object is represented very differently from other objects, VTC representations might still allow for appearance-invariant identification, e.g. to determine that a cow-shaped mug is still a mug. To allow this, the representation of a lookalike object should be more similar to another object from the same basic category (e.g., a cow-shaped mug and a regular mug) than to other objects. We operationalize this as the prediction that the within-triad correlation between each lookalike and its corresponding nonlookalike object would be significantly higher than the average of correlation of this same lookalike to objects from other triads. We call this the *identity index*. That is, for each subject and for each lookalike object, we took the on-diagonal correlation (e.g., between the cow-shaped mug and the mug) and subtracted the average of off-diagonal correlations (e.g., between the cow-shaped mug and the remaining objects). We computed the *identity index* separately for the animals (*animal identity index*: is a cow-shaped mug closer to a cow relative to other animals?) and the objects (*object identity index*: is a cow-shaped mug closer to a mug relative to other objects?). Figure 7 shows results for the full dataset. Supplementary Figure 3 reports results for the two tasks separately.

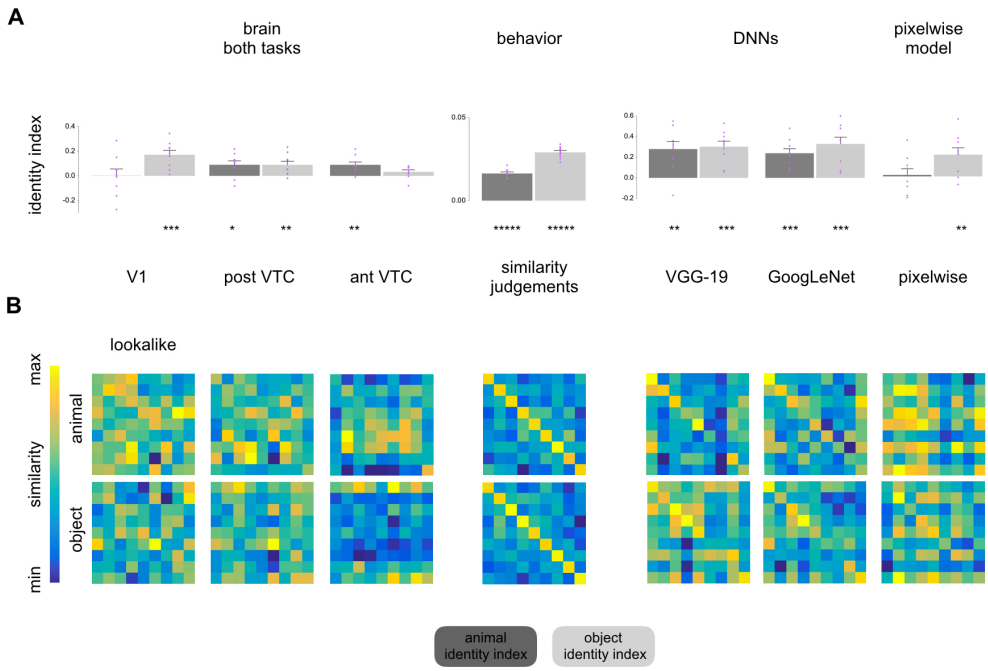


Figure 7. VTC representations do not sustain appearance-invariant object categorization **(A)** The *identity index* reflects information for individual stimulus pairs (e.g., cow-shaped mug and the mug) and is computed separately for each condition (animals and objects). For each lookalike object (n= 9) we took the on-diagonal correlation (e.g., between the cow-shaped-mug and the mug) and subtracted the average off-diagonal correlations (e.g., between the cow-shaped mug and the remaining objects). The identity index for animals and objects was computed for the brain data (V1, post VTC, and ant VTC), behavioral data (similarity judgments), DNNs (VGG-19, GoogLeNet), and the image pixelwise data. Light gray: animal identity index; dark gray: object identity index. Correlation matrices are shown for animals (and lookalikes), and objects (and lookalikes) separately. Asterisks (***** $p < 0.00001$, **** $< p < 0.0001$, *** $< p < 0.001$, ** $p < 0.01$, * $p < 0.05$) indicate significant values relative to baseline, and error bars indicate SEM. (B) For each dataset similarity matrices used to compute the identity index are shown separately for animals and objects.

Identity indexes were analyzed in a 2X3 ANOVA with Condition (animals, objects) and ROIs (V1, post VTC, ant VTC) as within-subject factors. Results revealed a significant Condition x ROI interaction ($F_{(2,16)}$ 11.47, $p = 0.001$; Figure 7A, left panel), thus revealing category-related differences in the amount of stimulus identity information carried in the three ROIs. Post-hoc analyses revealed that in V1 there was significant identity information for objects ($t_{(8)} = 4.77$, $p = 0.001$) but not for animals ($t < 1$). This result can be explained with differences in image low-level visual properties across conditions; objects and lookalikes were more similar to each other relative to animals and lookalikes. Confirming this interpretation, the same finding was observed for image pixelwise information (Figure 7A, right panel), where we observed significant above baseline identity information for objects ($t_{(8)} = 3.16$, $p = 0.01$) but not for animals ($t < 1$). While moving anterior into VTC results changed; object identity information decreased and animal identity information increased. Whereas the identity index was significantly above baseline for both conditions in posterior VTC (animals: $t_{(8)} = 2.82$, $p = 0.02$; objects $t_{(8)} = 3.09$, $p = 0.01$), in anterior VTC, only the animal identity index was significantly above baseline (animals: $t_{(8)} = 3.88$, $p = 0.005$; objects: $t_{(8)} = 1.93$, $p = 0.09$). Results were replicated when data were analyzed separately for the two tasks (Supplementary Figure 3). Together, these results suggest that representational content in anterior VTC is differently biased to represent animal and object's identity, containing more information for the former.

Different results were found in behavioral judgments and convolutional neural networks. The identity index performed on human similarity judgments revealed significant identity information for both animals ($t_{(8)} = 17.98$, $p < 0.00001$) and objects ($t_{(8)} = 24.44$, $p < 0.00001$; Figure 7A, middle panel). Similarly, DNNs were also able to discriminate individual stimulus identities for animals (VGG-19: $t_{(8)} = 3.84$, $p = 0.005$; GoogLeNet: $t_{(8)} = 5.42$, $p = 0.001$) as well as objects (VGG-19: $t_{(8)} = 5.67$, $p = 0.0005$; GoogLeNet: $t_{(8)} = 5.04$, $p = 0.001$; Figure 7A, right panel).

Strikingly, after years of searching for models that would match the tolerance in the most advanced category representations in primate ventral visual cortex, we have now found an object transformation for which neural networks sustain invariant object recognition to a higher degree than anterior VTC representations.

Discussion

A fundamental aim of cognitive computational neuroscience is to close the circle between behavior, brain, and machine computations. In other words, machines should behave like humans, and the behavior-brain and brain-machine computational correspondence should be well understood. With a stimulus set that specifically dissociates object appearance from object category we investigated whether dimensions underling the animacy organization are shared across ventral occipitotemporal cortex, deep neural networks, and behavior. We show that similarly to humans, neural networks can detect animacy. Conversely, representations in ventral occipitotemporal cortex mostly reflect object appearance. Thus, although deep neural networks can largely predict human behavior, representations in ventral occipitotemporal cortex deviate from behavior and neural network representations, indicating that the circle is yet to be closed.

Our results can be summarized with the following points. 1) Representational content in VTC reflects animal appearance as oppose to object identity and animacy. Thus, despite the mug and the cow-shaped mug share many high-level properties such as object identity, size, functional and manipulation knowledge, as well as low-level properties in our stimulus set, VTC represents a cow-shaped mug closer to a real cow than to a mug. 2) These neural data do not reflect either human perception or results from deep convolutional neural networks. Both humans and DNNs are not deceived by object appearance, and correctly judge a cow-shaped mug being closer to a mug than a real cow. 3) Finally, in VTC similar representations for animals and objects with animal appearance are not modulated by task demand, even when participants performed a task requiring focusing on object animacy. In what follows we will discuss these points in detail.

The animacy division (animate versus inanimate objects) is considered one of the main visual cortex organizational principles (e.g., Grill-Spector and Weiner, 2014), but which object properties underlie this and other organizational principles is highly debated (Baldassi et al., 2013; Bracci and Op de Beeck, 2016; Grill-Spector and Weiner, 2014; Kalfas et al., 2017; Nasr et al., 2014). Animacy and other category distinctions are often correlated with a range of low- and higher-level visual features such as the spatial frequency spectrum (Nasr et al., 2014; Rice et al., 2014) and shape (Cohen et al., 2014; Jozwik et al., 2016), but the animacy structure remains even when dissociated from such features (Bracci and Op de Beeck, 2016). However, here we question what we imply with animacy. It turns out that the dominant organizational principle is not about animate versus inanimate objects. The semantic and categorical meaning of animacy (what is the object and is it animate or not?) is close to irrelevant. Instead, what mostly matters is animal appearance: does the object look like an animate object or not?

We are not arguing that the animacy division in VTC is completely explained by animal appearance. Indeed, results from RSA showed that although the two models were uncorrelated ($r = 0.07$), we observed significant effects for both (Figure 4). Furthermore, results shown in Figure 6 (left panel) show that anterior VTC carries enough information to distinguish between real animals and lookalikes. What we argue is that information about animals in VTC is overrepresented relative to information about objects. To the extent that even inanimate objects, if having animal appearance, are represented similarly to animate entities.

A VTC bias towards animate representations was also supported by results showing significant information for animal's identity, which was not observed for objects (Figure 7). An everyday life example of this bias is our strong predisposition to perceive faces (Pareidolia) in meaningless patterns (e.g., toasts, clouds), which results in face-like activations in the fusiform face area (Hadjikhani et al., 2009). We could speculate that

having a system devoted to detect animate entities and biased to represent similarly animals and inanimate objects that resemble animals, might have been beneficial from an evolutionary perspective. The chance for our ancestors to survive in the wild necessitated fast and successful detection of predators. Such a biased system would have been advantageous: better mistaking a bush for a lion than the other way around. Furthermore, interactions with animate and inanimate objects require completely different skills; in the first case given their intention, ability to move, and react to our actions, a capacity to predict actions from either animals or conspecifics is a fundamental human skill that contributed to the evolution of our species. Thus, the bias of VTC to represent lookalikes similarly to real animals might be the result of an evolution-long "training", similarly to the way artificial intelligence neural networks, trained to recognize animals, start to detect animals in random-noise images (Szegedy et al., 2015).

This speculation points to one possible explanation for a particularly surprising aspect of our findings, namely the discrepancy between VTC representations at one hand and artificial DNN representation at the other hand. Here it is noteworthy that the training history of the DNNs was focused upon the categorization of a large number of categories and did not specifically bias towards e.g. animate objects. This training history might explain why DNNs are very good, and even better than human visual cortex, at identifying a cow-shaped mug as a mug.

To summarize, these results highlight substantial differences in the way categorical representations in the human brain, human judgments, and deep convolutional neural networks characterize objects. Whereas DNN representations resemble human behavior and heavily weigh object identity, representations in ventral visual cortex are very much constrained by object appearance. Based on these findings, we might have to reconsider the traditional view of human visual cortex as a general object recognition machine like the artificial neural networks are trained to be.

Methods

Participants

The study included 16 adult volunteers (9 males; mean age, 30 years). Informed consent to take part in the fMRI experiment was signed by all participants. The ethics committee of the KU Leuven approved the study. For the fMRI experiment, due to excessive head motion, all data from one participant was excluded. In addition, one run was excluded in two participants, and two runs were excluded in two other participants. The head motion exclusion criterion was set to +/- 3 mm (equal to 1 voxel size) and defined before data collection. For behavioral ratings, two participants were excluded due to technical problems during data collection.

Stimuli

The stimulus set included nine different triads, each containing: (1) one animal (e.g., a cow), (2) one object (e.g., a mug), and (3) one lookalike object, which resembled the animate one (e.g., a cow-shaped mug; Figure 1A). Critically, each lookalike object was carefully matched to the object, thus sharing the same object identity, size, and all object properties such as function and usage; the two mugs, apart from their shape, are very similar in all these other aspects. As for the animal condition, the lookalike stimulus represented the same animal (e.g., two cows) but differed in all aspects mentioned above (e.g., real-world size and function), but most importantly it differed in animacy; the cow-shape mug is obviously an inanimate object whereas the cow image depicts a living animate animal.

Behavioral Data

Each participant rated all images (Figure 1A) on object general similarity. Images were simultaneously rated on a screen within a circle arena following the procedure of Kriegeskorte and Mur (2012). Participants were asked to "arrange the objects according to how similar they are", thus letting them free to choose what they considered to be the most relevant dimension/s to judge. For each judgment (general similarity), results were averaged across participants and (as for the neural data) the upper triangle of the dissimilarity matrix was used for subsequent analyses.

fMRI Data

Experimental Design: We acquired neuroimaging data by means of an event-related design in two separated sessions, each performed in separated days with no more than 7 days between the first and second session. Each session included six experimental runs as well as additional runs with unrelated stimuli for another experiment (not reported here). The stimuli presentation was controlled by a PC running the Psychophysics Toolbox package (Brainard, 1997) in MATLAB (The MathWorks). Pictures were projected onto a screen and were viewed through a mirror mounted on the head coil.

Each experimental run (12 in total) lasted 7 min and 14 ms (230 volumes per run). For each subject and for each run, a fully randomized sequence of 27 image trials (repeated 4 times) and 9 fixation trials (repeated 4 times) was presented. Each trial was presented for 1500 ms, followed by a fixation screen for 1500 ms. Each run started and ended with 14 sec of fixation. Within the whole experiment, each stimulus was repeated 48 times. While scanning participants performed two different tasks (Figure 1B) counterbalanced across runs. During the animacy task (task 1) participants judged animacy ("does this image depict a living

animal?"). During the lookalike task (task 2) participants judged object appearance ("does this image look like an animal?"). Participants responded 'yes' or 'no' with the index and middle finger. Response-finger associations were counterbalanced across runs.

Acquisition Parameters: Imaging data was acquired on a 3T Philips scanner with a 32-channel coil at the Department of Radiology of the University Hospitals Leuven. MRI volumes were collected using echo planar (EPI) T2*-weighted scans. Acquisition parameters were as follows: repetition time (TR) of 2 s, echo time (TE) of 30 ms, flip angle (FA) of 90°, field of view (FoV) of 216 mm, and matrix size of 72 x 72. Each volume comprised 37 axial slices (covering the whole brain) with 3 mm thickness and no gap. The T1- weighted anatomical images were acquired with an MP-RAGE sequence, with 1x1x1 mm resolution.

Preprocessing: Imaging data were preprocessed and analyzed with the Statistical Parametrical Mapping software package (SPM 12, Welcome Department of Cognitive Neurology) and MATLAB. Before statistical analysis, functional images underwent a standard preprocessing procedure to align, co-register and normalize to an MNI (Montreal Neurological Institute) template. Functional images were spatially smoothed by convolution of a Gaussian kernel of 4 mm full width at half-maximum (Op de Beeck, 2010). For each participant, a general linear model (GLM) for both tasks was created to model the 27 conditions of interest and the 6 motion correction parameters (x, y, z for translation and for rotation). Each predictor's time course was modeled for 3 sec (stimulus presentation + fixation) by a boxcar function convolved with the canonical hemodynamic response function in SPM. We also analyzed data for each task separately, for which purpose a general linear model (GLM) was created for each task separately.

Regions of Interest (ROIs) definition: ROIs were defined at the group level with a combination of functional and anatomical criteria. First, we selected all visually active voxels (all stimuli versus baseline) that exceeded the statistical uncorrected threshold $p < 0.001$. Subsequently, we selected all spatially continuous voxels within anatomical ROIs defined with the Neuromorphometrics atlas in SPM. The following ROIs were defined: V1, and ventro-temporal cortex (VTC) divided into its posterior (post VTC) and anterior (ant VTC) sectors. The VTC ROIs were chosen based on the important role played by these areas in object representation and categorization (Grill-Spector and Weiner, 2014).

Statistical Analyses: Multivariate analyses were used to investigate the prevalence of object identity and object appearance in the representational content of ventral temporal cortex (VTC) under different task conditions. As before (Op de Beeck et al., 2010), for each voxel within a given ROI, parameter estimates for each condition (relative to baseline) were extracted for each participant and each run and normalized by subtracting the mean response across all conditions. Subsequently, the data set was divided 100 times into 2 random subsets of runs (set-1 and set-2) and the voxel response patterns for each object pair were correlated across these independent data sets. Correlations were averaged across the 100 iterations, thus resulting in an asymmetric 27 × 27 correlation matrix for each task, participant and ROI. For each correlation matrix, cells above and below the diagonal were averaged and only the upper triangle of the resulting symmetric matrix was used in the subsequent analyses (Ritchie et al., 2017). Correlation matrices were converted into dissimilarities matrices (1 minus Pearson's r) and used as neural input for subsequent analyses. Resulting correlations were

Fisher transformed {0.5*log[(1 + r)/(1 − r)]} and tested with ANOVAs and pairwise t-tests. Uncorrected p-values are reported.

For each ROI we computed an estimate of the reliability of the dissimilarity matrices (Op de Beeck et al., 2008). For each subject and each ROI, the 27 × 27 correlation matrix was correlated with the averaged correlation matrix of the remaining participants. The resulting correlation values (averaged across participants) capture noise inherent to a single subject as well as noise caused by inter-subject variability.

Deep Neural Networks

To create the object dissimilarity space of our stimulus set, we used the stimuli features that were extracted for the last processing stage (output layer) of two Deep Neural Networks (DNNs), VGG-19 (Simonyan and Zisserman, 2014) and GoogLeNet (Szegedy et al., 2015). These DNNs are deep convolutional neural networks, which have been very successful for object recognition tasks in the last few years. They consist of various processing stages, which are often termed as 'layers' either individually or in groups, in a feedforward manner that nonlinearly transform an input image volume (width, height, RGB value) into a 1-dimensional vector holding the class scores.

These processing stages include: i) convolutional layers, where a neuron outputs the dot product between a kernel (its weights) and a small region in the input it is connected to, ii) rectified linear unit activation functions, such that the activations of a previous weight layer are thresholded at zero (max(0,x)), iii) max pooling layers performing a downsampling operation along the width and height of the input and lastly, iv) fully connected layers (resembling a multilayer perceptron) that flatten the previous stage's volume and end up in a 1-dimensional vector with the same size as the number of classes. A softmax function is then typically applied to this last fully connected layer's unit activations to retrieve class probability scores for the classification task. In our experiments, we used the last fully connected layer as the upper layer of the network and not the probability scores from applying the softmax function.

The two networks were trained on ~1.2 million natural images, belonging to 1000 classes for the ImageNet Large Scale Visual Recognition Challenge 2012 (ILSVRC2012). The classes include ~40% animals and ~60% objects, but the networks are not specifically trained to detect animacy. We used the pretrained models from MatConvNet (http://www.vlfeat.org/matconvnet/)(Vedaldi and Lenc, 2016) toolbox in MATLAB. Before feature extraction, the mean of the ILSVRC2012 training images was subtracted from each image, as to the standards of training both these DNNs. Furthermore all stimuli were scaled to 224x224 pixels, in accordance with the requirements of the two networks.

VGG-19. This deep neural network from Simonyan and Zisserman (2014) incorporates 16 convolutional layers, in 2 blocks of 2 convolutional layers and 3 blocks of 4 convolutional layers, followed by three fully connected layers. A max pooling operation is applied after each of these five convolutional layer blocks. All, except from the last (fc8), of these weight layers are followed by a RELU function, where the activations are thresholded at zero. A softmax function is then applied to the last fully connected layer (fc8). The top-5 error rate performance of this pretrained MatConvNet model on the ILSVRC2012 validation data, was 9.9%.

GoogLeNet. Google's entry to ILSVRC2014 (Szegedy et al., 2015). It made use of the 'Inception' module, which is a technique used in early versions of Deep Neural Networks for pattern recognition, where a DNN uses several sizes of kernels along with pooling concatenated within one layer, which is similar to

integrating all information about parts of the image (size, location, texture etc.). In addition, there is a softmax operation in multiple stages of GoogLeNet, assisting the classification procedure during training along the depth levels of the network. This MatConvNet pretrained model was imported from the Princeton version, not by the Google team, thus there is some difference in performance to other versions probably due to parameter settings during training. The top-5 error rate performance of this model on the ILSVRC2012 validation data was 12.9%.

Data and software availability

All types of brain images and statistics data are available from the authors upon request. The data used for final statistics are made available through the Open Science Framework.

Author Contributions

S.B. and H.O.B designed research; S.B. performed research; S.B. and I.K. analyzed data; S.B., I.K. and H.O.B. wrote the paper.

Acknowledgements

This work was supported by the European Research Council (ERC- 2011-StG-284101), a federal research action (IUAP-P7/11), The KU Leuven Research Council (C14/16/031), and a Hercules grant ZW11_10 to H.O.P. S.B. was funded by FWO (Fonds Wetenschappelijk Onderzoek) through a postdoctoral fellowship (12S1317N).

References

Baldassi, C., Alemi-Neissi, A., Pagan, M., Dicarlo, J.J., Zecchina, R., and Zoccolan, D. (2013). Shape similarity, better than semantic membership, accounts for the structure of visual object representations in a population of monkey inferotemporal neurons. PLoS computational biology 9, e1003167.

Biederman, I. (1987). Recognition-by-components: a theory of human image understanding. Psychological review 94, 115-147.

Bracci, S., and Op de Beeck, H. (2016). Dissociations and Associations between Shape and Category Representations in the Two Visual Pathways. The Journal of neuroscience : the official journal of the Society for Neuroscience 36, 432-444.

Brainard, D.H. (1997). The Psychophysics Toolbox. Spatial vision 10, 433-436.

Cadieu, C.F., Hong, H., Yamins, D.L., Pinto, N., Ardila, D., Solomon, E.A., Majaj, N.J., and DiCarlo, J.J. (2014). Deep neural networks rival the representation of primate IT cortex for core visual object recognition. PLoS computational biology 10, e1003963.

Cichy, R.M., Khosla, A., Pantazis, D., Torralba, A., and Oliva, A. (2016). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. Scientific reports 6:27755.

Cohen, M.A., Konkle, T., Rhee, J.Y., Nakayama, K., and Alvarez, G.A. (2014). Processing multiple visual objects is limited by overlap in neural channels. Proceedings of the National Academy of Sciences of the United States of America 111, 8955-8960.

Desimone, R., Albright, T.D., Gross, C.G., and Bruce, C. (1984). Stimulus-selective properties of inferior temporal neurons in the macaque. The Journal of neuroscience : the official journal of the Society for Neuroscience 4, 2051-2062.

Grill-Spector, K., Kushnir, T., Edelman, S., Avidan, G., Itzchak, Y., and Malach, R. (1999). Differential processing of objects under various viewing conditions in the human lateral occipital complex. Neuron 24, 187-203.

Grill-Spector, K., Kushnir, T., Edelman, S., Itzchak, Y., and Malach, R. (1998). Cue-invariant activation in object-related areas of the human occipital lobe. Neuron 21, 191-202.

Grill-Spector, K., and Weiner, K.S. (2014). The functional architecture of the ventral temporal cortex and its role in categorization. Nature reviews Neuroscience 15, 536-548.

Gross, C.G., Rocha-Miranda, C.E., and Bender, D.B. (1972). Visual properties of neurons in inferotemporal cortex of the Macaque. Journal of neurophysiology 35, 96-111.

Guclu, U., and van Gerven, M.A. (2014). Unsupervised feature learning improves prediction of human brain activity in response to natural images. PLoS computational biology 10, e1003724.

Guclu, U., and van Gerven, M.A. (2015). Deep Neural Networks Reveal a Gradient in the Complexity of Neural Representations across the Ventral Stream. The Journal of neuroscience : the official journal of the Society for Neuroscience 35(27):10005-10014.

Hadjikhani, N., Kveraga, K., Naik, P., and Ahlfors, S.P. (2009). Early (M170) activation of face-specific cortex by face-like objects. Neuroreport 20, 403-407.

Haxby, J.V., Gobbini, M.I., Furey, M.L., Ishai, A., Schouten, J.L., and Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science 293, 2425-2430.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. arXiv 1502.01852v1.

Jozwik, K.M., Kriegeskorte, N., and Mur, M. (2016). Visual features as stepping stones toward semantics: Explaining object similarity in IT and perception with non-negative least squares. Neuropsychologia 83, 201-226.

Kalfas, I., Kumar, S., and Vogels, R. (2017). Shape Selectivity of Middle Superior Temporal Sulcus Body Patch Neurons. eNeuro 4.

Kanwisher, N., McDermott, J., and Chun, M.M. (1997). The fusiform face area: a module in human extrastriate cortex specialized for face perception. The Journal of neuroscience : the official journal of the Society for Neuroscience 17, 4302-4311.

Khaligh-Razavi, S.M., and Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain IT cortical representation. PLoS computational biology 10, e1003915.

Khaligh-Razavi, S.M., Henriksson, L., Kay, K., and Kriegeskorte, N. (2014). Explaining the hierarchy of visual representational geometries by remixing of features from many computational vision models. bioRxiv doi: http://dx.doi.org/10.1101/009936.

Kheradpisheh, S.R., Ghodrati, M., Ganjtabesh, M., and Masquelier, T. (2016a). Deep Networks Can Resemble Human Feed-forward Vision in Invariant Object Recognition. Scientific reports 6, 32672.

Kheradpisheh, S.R., Ghodrati, M., Ganjtabesh, M., and Masquelier, T. (2016b). Humans and Deep Networks Largely Agree on Which Kinds of Variation Make Object Recognition Harder. Frontiers in computational neuroscience 10, 92.

Kiani, R., Esteky, H., Mirpour, K., and Tanaka, K. (2007). Object category structure in response patterns of neuronal population in monkey inferior temporal cortex. Journal of neurophysiology 97, 4296-4309.

Konkle, T., and Oliva, A. (2012). A real-world size organization of object responses in occipitotemporal cortex. Neuron 74, 1114-1124.

Kriegeskorte, N., and Kievit, R.A. (2013). Representational geometry: integrating cognition, computation, and the brain. Trends in cognitive sciences 17, 401-412.

Kriegeskorte, N., and Mur, M. (2012). Inverse MDS: Inferring Dissimilarity Structure from Multiple Item Arrangements. Frontiers in psychology 3, 245.

Kriegeskorte, N., Mur, M., and Bandettini, P. (2008a). Representational similarity analysis - connecting the branches of systems neuroscience. Frontiers in systems neuroscience 2, 4.

Kriegeskorte, N., Mur, M., Ruff, D.A., Kiani, R., Bodurka, J., Esteky, H., Tanaka, K., and Bandettini, P.A. (2008b). Matching categorical object representations in inferior temporal cortex of man and monkey. Neuron 60, 1126-1141.

Kriegeskorte, N. (2015). Deep Neural Networks: A New Framework for Modeling Biological Vision and Brain Information Processing. Annual review of vision science 1:417-446.

Krizhevsky A., Sutskever I., Hinton G.E (2012). ImageNet classification with deep convolutional neural networks. In: Pereira F, Burges CJC, Bottou L, Weinberger KQ, editors. Advances in Neural Information Processing Systems 1097–105.

Kruschke, J.K. (1992). ALCOVE: an exemplar-based connectionist model of category learning. Psychological review 99, 22-44.

Kubilius, J., Bracci, S., and Op de Beeck, H.P. (2016). Deep Neural Networks as a Computational Model for Human Shape Sensitivity. PLoS computational biology 12, e1004896.

Martin, A., Wiggs, C.L., Ungerleider, L.G., and Haxby, J.V. (1996). Neural correlates of category-specific knowledge. Nature 379, 649-652.

Medin, L.D., and Schaffer, M.M. (1978). Context Theory of Classification Learning. Psychological review 85, 207-238.

Nasr, S., Echavarria, C.E., and Tootell, R.B. (2014). Thinking outside the box: rectilinear shapes selectively activate scene-selective cortex. The Journal of neuroscience : the official journal of the Society for Neuroscience 34, 6721-6735.

Nosofsky, R.M. (1984). Choice, similarity, and the context theory of classification. Journal of experimental psychology Learning, memory, and cognition 10, 104-114.

Op de Beeck, H., Wagemans, J., and Vogels, R. (2001). Inferotemporal neurons represent low-dimensional configurations of parameterized shapes. Nature neuroscience 4, 1244-1252.

Op de Beeck, H.P. (2010). Against hyperacuity in brain reading: spatial smoothing does not hurt multivariate fMRI analyses? NeuroImage 49, 1943-1948.

Op de Beeck, H.P., Brants, M., Baeck, A., and Wagemans, J. (2010). Distributed subordinate specificity for bodies, faces, and buildings in human ventral visual cortex. NeuroImage 49, 3414-3425.

Op de Beeck, H.P., Torfs, K., and Wagemans, J. (2008). Perceived shape similarity among unfamiliar objects and the organization of the human object vision pathway. The Journal of neuroscience : the official journal of the Society for Neuroscience 28, 10111-10123.

Rice, G.E., Watson, D.M., Hartley, T., and Andrews, T.J. (2014). Low-level image properties of visual objects predict patterns of neural response across category-selective regions of the ventral visual pathway. The Journal of neuroscience : the official journal of the Society for Neuroscience 34, 8837-8844.

Ritchie, J.B., Bracci, S., and Op de Beeck, H. (2017). Avoiding illusory effects in representational similarity analysis: What (not) to do with the diagonal. NeuroImage 148, 197-200.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. (2014). ImageNet Large Scale Visual Recognition Challenge. arXiv 1409.0575v3.

Sha, L., Haxby, J.V., Abdi, H., Guntupalli, J.S., Oosterhof, N.N., Halchenko, Y.O., and Connolly, A.C. (2015). The animacy continuum in the human ventral vision pathway. Journal of cognitive neuroscience 27, 665-678.

Simonyan, K., and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv 1409.1556.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going Deeper With Convolutions. arXiv 1409.4842v1

Tarr, M.J., and Bulthoff, H.H. (1998). Image-based object recognition in man, monkey and machine. Cognition 67, 1-20.

Vedaldi, A., and Lenc, K. (2016). MatConvNet: Convolutional Neural Networks for MATLAB. arXiv:14124564v3.

Xia, M., Wang, J., and He, Y. (2013). BrainNet Viewer: a network visualization tool for human brain connectomics. PloS one 8, e68910.

Yamins, D.L., Hong, H., Cadieu, C.F., Solomon, E.A., Seibert, D., and DiCarlo, J.J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. Proceedings of the National Academy of Sciences of the United States of America 111, 8619-8624.

Figure legends

Figure 1 The experimental design. (A) The stimulus set was specifically designed to dissociate object appearance from object identity. We included 9 different object triads. Each triad included an animal (e.g., butterfly), an inanimate object (e.g., earring), and a lookalike object closely matched to the inanimate object in terms of object identity and to the living animal in terms of object appearance (e.g., a butterfly-shaped earring). (B) During fMRI acquisition, participants performed two different tasks counterbalanced across runs. During the *animacy task* (task 1) participants judged animacy ("does this image depict a living animal?"). During the *lookalike task* (task 2) participants judged animal appearance ("does this image look like an animal?"). Participants responded 'yes' or 'no' with the index and middle finger. Responses were counterbalanced across runs. (C) Model predictions based on the two tasks. The animacy model predicts high similarity between the object and the lookalike condition, thus predicting all inanimate objects (objects and lookalikes) to cluster separately from living animals. On the contrary, the lookalike model predicts high similarities between the animal and the lookalike condition despite differences in object identity, functional properties, size, and animacy. The two models are independent ($r = 0.07$).

Figure 2 Similarities and differences between DNNs representations and human judgments. Representational similarity analysis (RSA; Kriegeskorte et al., 2008a) results for the lookalike model (dark gray) and the animacy model (light gray) are shown for human judgments (left panel), and DNNs (right panel). For behavioral results, individual participant's correlation values (n=15) are shown in purple. Asterisks indicate significant values relative to baseline, and error bars indicate SEM. For DNNs results, asterisks indicate significant values computed with permutation tests (10,000 randomizations of stimulus labels), and error bars indicate SE computed by bootstrap resampling of the stimuli. ***** $p < 0.00001$, **** $p < 0.0001$, *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

Figure 3 Deep neural networks and human perception predict a representation based on object animacy. (A-B) Dissimilarities matrices (upper panel) and 2D arrangements derived from MDS (metric stress; lower panel) show stimuli pairwise distances for behavioral judgments (A) and DNNs (B). Light blue: animals; blue: lookalikes; dark blue: objects.

Figure 4 Animal lookalike explains representational content in human visual cortex. (A) Group averaged ROIs are shown on an inflated human brain template in BrainNet Viewer (Xia et al., 2013). (B-D) Representational similarity analysis (RSA; Kriegeskorte et al., 2008) results for the lookalike model (dark gray) and the animacy model (light gray) are shown for the data combined across the two tasks (B), and for each task separately (C-D). Individual participant's (n =16) correlation values are shown in purple. Purple-shaded backgrounds represent reliability values of the correlational patterns taking into account the noise in the data. These values give an estimate of the highest correlation we can expect in each ROI. Error bars indicate SEM. Asterisks indicate significant difference between the two models (*** $p < 0.001$; ** $p < 0.01$).

Figure 5 The neural similarity space in VTC reflects animal appearance. (A) The neural dissimilarity matrices (1 minus Person's r) derived from the neural data (averaged across subjects and tasks) show pairwise dissimilarities among stimuli in the three ROIs. (B) The MDS (metric stress), performed on the dissimilarity matrices averaged across subjects and tasks, shows pairwise distances in a 2D space for the three ROIs. Animals: light blue; lookalike: blue; and objects: dark blue.

Figure 6 Differential category discriminability information in VTC and DNNs. The category index reflects representational discriminability among the three stimulus categories (animals, lookalikes, and objects) and is computed for each condition pair by subtracting the average of between-condition correlations (e.g., for animals and lookalikes) from the average of within-condition correlations (e.g., for animals and lookalikes). Results are reported for neural data (left panel), behavior (middle panel), and DNNs (right panel). Light gray: animals-lookalikes; gray: lookalikes-objects; dark gray: animals-objects. For neural and behavioral data, individual participant's (n=16) data points are shown in purple. Asterisks indicate significant values relative to baseline, and error bars indicate SEM. For DNNs data, asterisks indicate significant values computed with

permutation tests (10,000 randomizations of stimulus labels), and error bars indicate SE computed by bootstrap resampling of the stimuli. ***** $p < 0.00001$, **** $< p < 0.0001$, *** $< p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

Figure 7 VTC representations do not sustain appearance-invariant object categorization **(A)** The *identity index* reflects information for individual stimulus pairs (e.g., cow-shaped mug and the mug) and is computed separately for each condition (animals and objects). For each lookalike object (n= 9) we took the on-diagonal correlation (e.g., between the cow-shaped-mug and the mug) and subtracted the average off-diagonal correlations (e.g., between the cow-shaped mug and the remaining objects). The identity index for animals and objects was computed for the brain data (V1, post VTC, and ant VTC), behavioral data (similarity judgments), DNNs (VGG-19, GoogLeNet), and the image pixelwise data. Light gray: animal identity index; dark gray: object identity index. Correlation matrices are shown for animals (and lookalikes), and objects (and lookalikes) separately. Asterisks (***** $p < 0.00001$, **** $< p < 0.0001$, *** $< p < 0.001$, ** $p < 0.01$, * $p < 0.05$) indicate significant values relative to baseline, and error bars indicate SEM. (B) For each dataset similarity matrices used to compute the identity index are shown separately for animals and objects.

Supplementary Results

Task-specific ROI-based RSA

To investigate any task-related modulation, separately for each task, correlation values were tested in a 3x2 ANOVA with ROI (V1, VTC-post, VTC-ant) and Model (lookalike, animacy) as within-subject factors. No differences were observed relative to results from the full dataset (Figure 4C-D). In both cases, results revealed a significant ROI x Model interaction (task 1: $F_{(2,30)}$ 5.66, $p = 0.008$; task 2: $F_{(2,30)}$ 5.28, $p = 0.01$). Further confirming results reported in the main text, for both tasks, the neural data was significantly more correlated with the lookalike model than the animacy model in posterior VTC (task 1: $t_{(15)} = 3.83$, $p = 0.002$; task 2: $t_{(15)} = 3.05$, $p = 0.008$) and anterior VTC (task 1: $t_{(15)} = 2.97$, $p = 0.009$; task 2: $t_{(15)} = 3.44$, $p = 0.004$). Furthermore, in anterior VTC correlations with both models were significant for each task: lookalike model (task 1: $t_{(15)} = 7.60$, $p < 0.00001$; task 2: $t_{(15)} = 6.68$, $p < 0.00001$); animacy model (task 1: $t_{(15)} = 5.63$, $p = 0.00005$; task 2: $t_{(15)} = 5.03$, $p = 0.0001$). Instead, in posterior VTC, only correlations with the lookalike model significantly differed from baseline (task 1: $t_{(15)} = 4.77$, $p = 0.0002$; task 2: $t_{(15)} = 4.23$, $p = 0.0006$). Here, correlations with the animacy model where either not significant in task 1($t < 1.5$) or only marginally significant in task 2 ($t_{(15)} = 2.32$, $p = 0.03$). Finally, in V1 neither model in either task was significantly correlated with the neural data ($t < 1.5$, for both tasks and models).

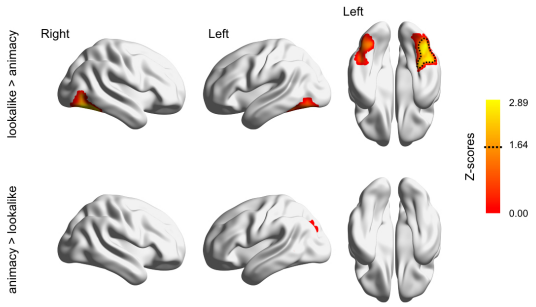
To further investigate potential task effects, separately for each ROI, correlation values were analyzed in a 2X2 ANOVA with Task (task 1, task 2) and Model (lookalike, animacy) as within-subject factors. In both VTC ROIs, results confirmed lack of task-related effects, revealing a main effect of Model (ant VTC: $F_{(1,15)}$ 13.71, $p = 0.002$; post VTC: $F_{(1,15)}$ 14.06, $p = 0.002$) and no effect of Task (VTC-ant: $F < 1$; VTC-post: $F < 1$).

Whole-brain RSA

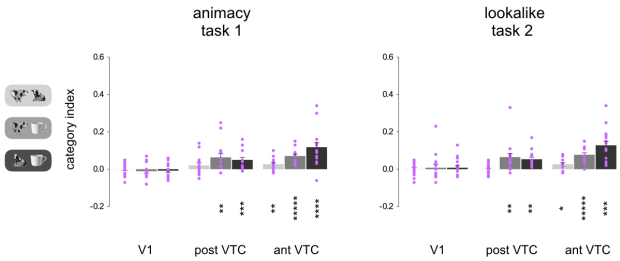
Do brain areas aside VTC show differential lookalike and animacy effects? To investigate this we performed a whole-brain RSA correlating the two models (lookalike, animacy) with neural patterns throughout the brain. The whole-brain RSA performed using the volume-based searching approach (Kriegeskorte et al., 2006) was implemented with CoSMo MVPA (Oosterhof et al., 2016). Parameter estimates for each condition (relative to baseline) were extracted for each participant and each run and normalized by subtracting the mean response across all conditions. Resulting values were then averaged across all runs. For each voxel in the brain a searchlight was defined using a spherical neighborhood with a variable radius, including the 100 voxels nearest to the center voxel. For each searchlight the neural dissimilarity matrix was computed for the 27 stimuli. The neural dissimilarity matrix (upper triangle) was then correlated with the dissimilarity matrices derived from the two predictive models (Figure 1C). The output correlation values were Fisher transformed and assigned to the center voxel of the sphere. Resulting whole-brain correlation maps for each of the models were directly contrasted and differences were tested using random-effects whole-brain group analysis and corrected with the Threshold-Free Cluster Enhancement (TFCE; Smith and Nichols, 2009) method. Voxel-wise corrected statistical maps ($z = 1.64$; $p < 0.05$, one-sided t-test) are highlighted with dotted lines were displayed on a brain template by means of BrainNet Viewer (Xia et al., 2013) .

Results confirmed conclusions from ROI-based RSA analysis. The direct contrast between the lookalike model and the animacy model (lookalike > animacy) revealed a significant cluster in (right) VTC ($p < 0.05$, one-

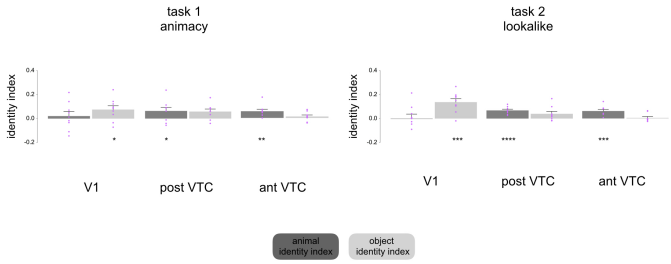
tail t-test, TFCE corrected). The opposite contrast (animacy > lookalike) revealed a non-significant trend in the occipito-parietal cortex.



Supplementary Figure 1 Whole-brain representational similarity analysis. Results of random-effects whole-brain RSA, for the direct contrast between the two predictive models. All colored voxels survived the voxel-wise Threshold-Free Cluster Enhancement correction (TFCE) with Monte Carlo simulation (n = 10000). For visualization purposes statistical maps are shown without threshold. Statistical maps surviving p < 0.05 (one-sided t-test) are highlighted with dotted lines.



Supplementary Figure 2 Category index for the animacy and the lookalike task. The category index reflects representational discriminability among the three stimulus categories (animals, lookalikes, and objects) and is computed for each condition pair by subtracting the average of between-condition correlations (e.g., for animals and lookalikes) from the average of within-condition correlations (e.g., for animals and lookalikes). Results for neural data are reported for the two tasks separately. Light gray: animals-lookalikes; gray: lookalikes-objects; dark gray: animals-objects. Individual participant's (n=16) data points are shown in purple. Asterisks indicate significance values relative to baseline (***** p < 0.00001, **** < p < 0.0001, *** < p < 0.001, ** p < 0.01, * p < 0.05), and error bars indicate SEM.



Supplementary Figure 3 Identity index for the animacy and lookalike task. The *identity index* reflects information for individual stimulus pairs (e.g., cow-shaped mug and the mug) and is computed separately for each condition (animals and objects). For each lookalike object (n = 9) we took the on-diagonal correlation (e.g., between the cow-shaped-mug and the mug) and subtracted the average off-diagonal correlations (e.g., between the cow-shaped mug and the remaining objects). The identity index for animals and objects was computed for the brain data (V1, post VTC, and ant VTC), separately for the animacy task (left panel), and the lookalike task (right panel). Light gray: animal identity index; dark gray: object identity index. Asterisks (**** p < 0.0001, *** < p < 0.001, ** p < 0.01, * p < 0.05) indicate significance values relative to baseline and error bars indicate SEM. Results were not modulated by task.

References

Kriegeskorte, N., Goebel, R., and Bandettini, P. (2006). Information-based functional brain mapping. Proceedings of the National Academy of Sciences of the United States of America 103(10):3863-3868.

Oosterhof, N.N., Connolly, A.C., and Haxby, J.V. (2016). CoSMoMVPA: Multi-Modal Multivariate Pattern Analysis of Neuroimaging Data in Matlab/GNU Octave. Frontiers in neuroinformatics 10:27.

Smith, S.M. and Nichols, T.E. (2009). Threshold-free cluster enhancement: addressing problems of smoothing, threshold dependence and localisation in cluster inference. NeuroImage 44(1):83-98.

Xia, M., Wang, J., and He, Y. (2013). BrainNet Viewer: a network visualization tool for human brain connectomics. PloS one 8(7):e68910.