Title: Body size as a metric to define objects in an affordable world
Running title: body-scaled metric for affordance.

Xinran Feng[1†], Shan Xu[2†], Yuannan Li[1], & Jia Liu[1*]

[1]Department of Psychology & Tsinghua Laboratory of Brain and Intelligence, Tsinghua University, Beijing, China.
[2]Faculty of Psychology, Beijing Normal University, Beijing, China
[†]Equal contribution
[*]Correspondence to: liujiathu@tsinghua.edu.cn (J. Liu)

Abstract

The physical body of an organism serve as a vital interface for interactions between the organism and its environment, significantly shaping its intelligence. Here we investigated the impact of human body size on the perception of affordance that is the action possibilities offered by environment to humans. Our findings revealed that body size delineated a distinct boundary on affordance similarity, dividing objects with continuous real-world sizes into two discrete categories: objects on either side of the boundary afford distinct sets of affordances. In addition, the boundary adjusted in accordance with changes in imagined body size, indicating a close link between body size and affordance perception. Intriguingly, the boundary may not be exclusively derived from organism-environment interactions, as ChatGPT, a large language model lacking a physical embodiment, exhibited a modest yet comparable affordance boundary at the scale of human body size. This implies that a human-like body schema may emerge in ChatGPT through exposure to linguistic materials alone. A subsequent fMRI experiment explored the functionality of the boundary, determining that only the affordances of objects within the range of body size were represented in the visual streams of the brain. This suggests that objects capable of being manipulable are the only objects capable of offering affordance in the eyes of the organism. In summary, our study presents an embodied perspective on defining object-ness in an affordable world, advocating

the concept of embodied cognition for understanding the emergence of intelligence under the constraints of an organism's physical attributes.

Keywords: embodied cognition; affordance; body-scaled information; affordance boundary; ChatGPT;

Introduction

Man is the measure of all things. - Protagoras

The assertion by the ancient Greek philosopher Protagoras highlights the notion that reality is defined by how the world is perceived by humans. A contemporary interpretation of this statement is the embodied theory of cognition (e.g., Chemero, 2013; Gallagher, 2017; Wilson, 2002; Varela et al., 2017), which posits that humans body scale (e.g., size, shape) shape object perception and generation of motor responses, encompassing locomotion and object manipulations. For instance, humans assess the climbability of steps based on their leg length (Mark, 1987; Warren,1984), and ascertain whether they can traverse an aperture according to the critical aperture-to-shoulder-width ratio (Warren & Whang, 1987). Additionally, the way of grasping has been shown to be contingent upon object size relative to body size (Cesari & Newell, 2000; Newell et al., 1989) or hand size (Castiello et al., 1993; Tucker & Ellis, 2004). However, it is unclear whether, besides as a reference for locomotion and object manipulation, body size plays a pivotal role in shaping the representation of objects, as suggested by Protagoras.

To underscore the latter point, Gibson (1979) posited that "Detached objects must be comparable in size to the animal under consideration if they are to afford behavior (p.124)." That is to say, an object's affordance, which encompasses all action possibilities that the object offer to an animal, is determined by the object's size relative to the animal's size, rather than its real-world size wherein objects are assessed independently of the animal's size.

Specifically, there may exist a qualitative distinction between objects within and beyond the size range of the animal, with only those objects within the animal's size affording interactions. Consequently, the body likely serves as a metric that facilitates meaningful engagement with the world by deciding which objects are accessible for interaction. Based on Gibson's intuition, we proposed two hypotheses: first, the affordance of objects will exhibit a qualitative difference between objects within and beyond the size range of an organism's body; second, affordance-related cortical regions (e.g., Roux-Sibilon et al., 2018) will be engaged exclusively by objects within the organism's size range.

To test these hypotheses, we first measured the affordance of a diverse array of objects varying in real-world size (e.g., Konkle, & Oliva, 2011). We found a dramatic decline in affordance similarity between objects within and beyond human body size range, as they afforded distinct sets of action possibilities. Specifically, the affordance boundary shifted in accordance with imaginary body sizes, and could be attained solely from language as demonstrated by the large language model (LLM), ChatGPT (OpenAI, San Francisco, CA, USA, 2022). A subsequent fMRI experiment corroborated the qualitative difference in affordance demarcated by the body size, as the affordance of objects within humans' size range, but not those beyond, was represented in both dorsal and ventral visual streams of the brain.

Results

To illustrate how agent size affected the affordance of object with different sizes, we first characterized the affordance of a set of daily objects. Specifically, in each trial we presented an image matrix consisting of nine objects and asked participants to report which objects afford a specific action (e.g., sit-able: a chair, a bed, a ladder, and a bike, but not an apple, a plate, a potted plant, a phone, and a hammer) (Figure 1A). In this task, there were 14 actions commonly executed in daily life and 24 object images from the THINGS database (Hebart et al., 2019) with the size ranging from size rank 2 to 8 in Konkle and Oliva (2011)'s

classification, covering sizes from much smaller (14 cm, rank 2) than the human body to those dozens of times larger (7618 cm, rank 8) than the human body in real-world (see Methods for details). The affordance for each object was indexed by a 14-dimension action vector, and the value for each dimension is the percentage of the participants agreed on certain action affordable by the object (e.g., 88% was the value for the dimension of grasping on a hammer because 88% of participants agreed that a hammer can afford grasping). Figure 1B shows the affordance of two example objects.

An affordance similarity matrix was then constructed where each cell corresponded to the similarity in affordances between a pair of objects (Figure 1C). A clustering analysis revealed a two-cluster structure, with the upper-left cluster consisting of objects whose size was smaller than the human body size (labeled with red shades), and the lower-right containing objects whose size was beyond the human body size (labeled with blue shades). Critically, the between-cluster similarity in the affordance similarity matrix was close to zero, suggesting a divide between the affordance separated by body size. To clearly illustrate this, we calculated the similarity in affordance between each neighboring size rank pair. Indeed, we observed a clear trough in affordance similarity, down to around zero, between size rank 4 (77cm on average) and 5 (146cm on average), which was significantly smaller than that between size rank 3 and 4 ($Z = 3.91$, $p < .001$) and that between size rank 5 and 6 ($Z = 1.66$, $p = .048$). Except this point, the rest affordance similarity stayed high (rs > 0.45) and did not significant differ from each other (ps > 0.05) (Figure 1D, blue line). Note that this abrupt change in affordance similarity across the boundary cannot be explained by changes in object size itself, as the similarity in object size remained relatively unchanged, without any trend of a trough shape curve (Figure 1D, grey line). Intriguingly, the size that the boundary locates corresponded to the size of a typical human adult (in size rank 5), suggesting that based on their affordance, objects were classified into two categories bordering exactly at the body size of humans.

An analysis on the affordances separated by the boundary shows that the objects on the left (within the body size) mainly afforded hand-related actions, such as grasp, hold, and throw. In contrast, the objects on the right (beyond the body size) did not afford this set of

action; instead, they mainly afforded actions such as sit, stand, push, and lift, which require the action of whole body. In addition, the affordance of the objects on the left tended to be more selective, converging on a smaller number of actions, than those on the right (Supplementary Figure S2, left). Consistent with this observation, the average affordance similarities were higher (mean r = 0.95) between size ranks within the body size than those beyond (mean r = 0.52). To quantify the difference in affordance, we measured the sparseness of affordance on either side of the boundary ($sparseness_{beyond}$ = 0.77, $sparseness_{within}$ = 0.56, see Methods), with the difference being statistically significant (permutation N = 5000, $p(T > T_{obs})$ = 0.024), suggesting that the affordance of the objects beyond was less selective than those within.

To establish a causal link between the affordance boundary and the body size, we directly manipulated the body schema (i.e., body representation) (Merleau-Ponty & Smith, 1962) to examine whether the affordance boundary shifted accordingly. With the same paradigm (Figure 1A), we instructed a new group of participants to imagine themselves to be as small as a cat (in size rank 4, 77cm on average, the cat condition), and another new group of participants an elephant (in size rank 7, 577 cm on average, the elephant condition) during the whole task (Figure 1E). This manipulation was successful, as at the end of the task the participants reported imaginary height in the cat condition being 42 cm and 450 cm in the elephant condition on average, respectively.

Importantly, with exactly the same set of objects, the affordance showed a clear boundary shift for each condition (Figure 1F). The affordance boundary in the cat condition was between size rank 3 and 4, where the similarity of size rank 3-4 was significantly lower than that of size rank 2-3 (Z = 1.76, p = .039) and that of 4-5 (Z = 1.68, p = .047). In contrast, in the elephant condition a sharp drop between the similarity of size rank 5-6 and those of size rank 6-7 and size rank 7-8 was observed, with a trend in significance (with size rank 6-7: Z = 1.28, p = .099; with size rank 7-8: Z = 1.48, p = .069). The distinction in sparseness of the affordance between the object within the imaginary body size and beyond showed a similar pattern when the body schema was directly manipulated (Supplementary Figure S2, right).

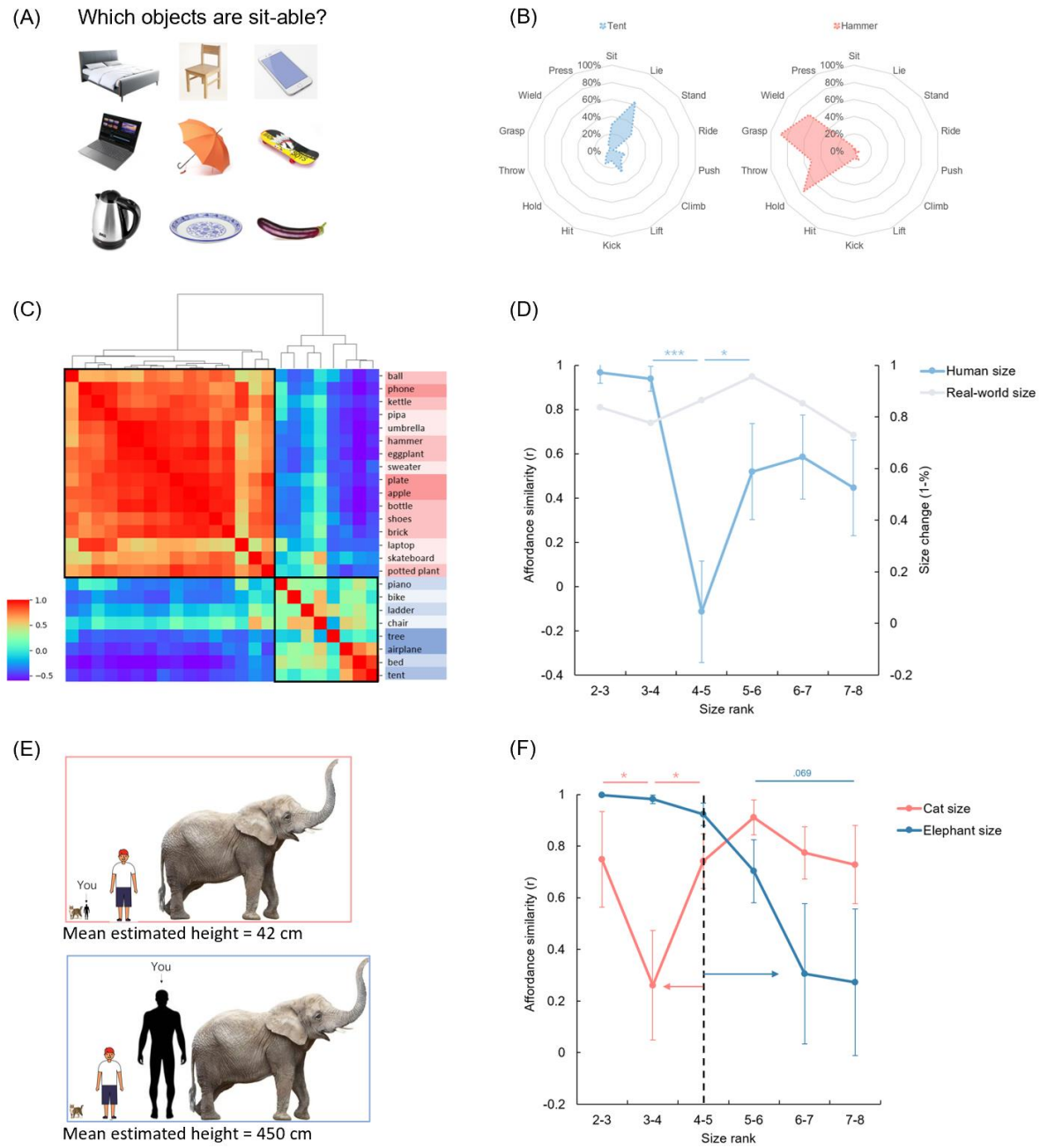

Figure 1. (A) An illustration of the object-action relation rating task. (B) Two exemplars of objects with different profiles of action possibilities. The values represented the probabilities of participants considering the given object affording each action. (C) The similarity matrix of objects based on affordance similarity. Warm colours in each voxel indicated high similarity. Object names were shaded according to object size (red-white-blue: small to big, with shade saturation corresponding with the exact sizes). The two primary clusters were marked with black boxes. (D) The affordance similarity (blue line; left y-axis) and size similarity (grey line; right y-axis) between neighboring size ranks. (E) The two

introduction images used in the cat condition (top panel) and the elephant condition (bottom panel), with the mean estimated height reported by participants shown separately. (F) The affordance similarity between neighboring size ranks with manipulated agent sizes (red line: cat-size agent; blue line: elephant-size agent). The dashed line represented the boundary of the human-size agent. The red and blue arrows indicated the corresponding boundary shift in each agent-size condition.

Where did the affordance boundary originate? A representative embodied view, the enactive theory, argues that affordance only originates from interactions between animals and their environment (Gallagher, 2017; Thompson, 2010). However, it has been proposed that conceptual knowledge are grounded in the same neural systems as those involved in perception and action (Barsalou, 1999), and therefore the sensorimotor information underlying the interactions with the environment is already embedded in language (e.g., Casasanto, 2011; Glenberg & Gallese, 2012; Stanfield & Zwaan, 2001). Therefore, an alternative possibility is that an organism may gain the affordance through language materials only. To test these two alternatives, here we employed a bodyless agent, a large language model (LLM) of ChatGPT (Chat Generative Pre-trained Transformer; https://openai.com/blog/chatgpt/), which has been trained on massive corpus of language materials derived from human and yet receive no sensorimotor information from the environment. Here we asked whether language alone is sufficient to form the affordance boundary in ChatGPT, as well as in smaller LLMs BERT and GPT-2.

The experimental procedure was similar to that conducted by human participants except that images were replaced by the corresponding words. Given randomness embedded in the model, each model was tested 20 times to mimic the sampling of human participants. We found that the affordance similarity curve demonstrated by the ChatGPT was trough-shaped between size rank 4 and 5, the same location where the boundary was found in human participants (Figure 2, red line). Further statistical analyses showed a significant difference in affordance similarity between size 3-4 and size 4-5 ($Z = 1.98$, $p = .024$), and the difference between size 4-5 and size 5-6 only showed a trend but did not reach the significance (Z =

0.96, p = .17). In contrast, there was no trough-shaped boundary observed in either BERT or GPT-2 (Figure 2, green line for BERT, yellow line for GPT-2). Though there was an apparent drop in affordance similarity in GPT-2 between size rank 5-6, the difference did not reach significance (p > .20). Apparently, the extent to which the performance resemblance in affordance perception of the LLMs with human participants increased with the size of models (BERT < GPT-2 < ChatGPT).

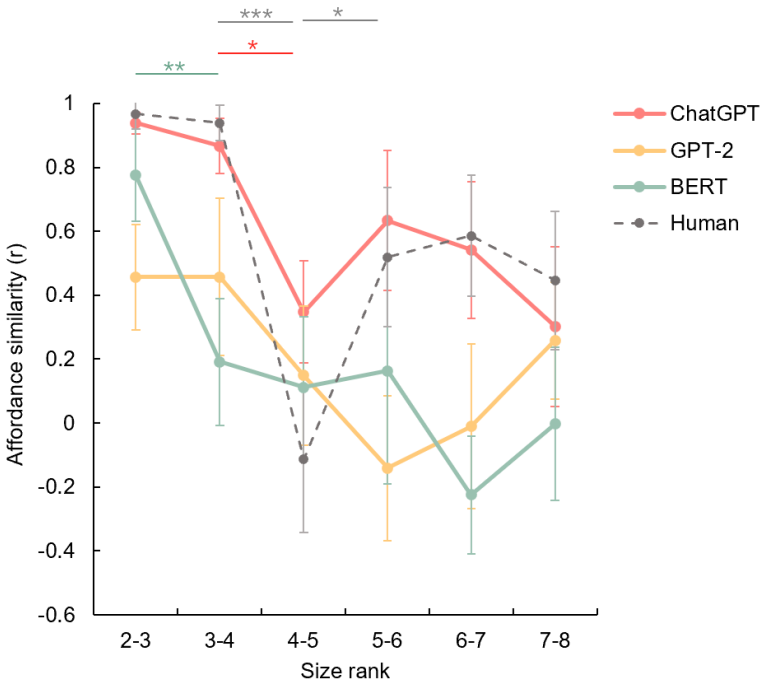

Figure 2. The affordance similarity between neighboring size ranks with the outputs from different large language models (red line: ChatGPT; yellow line: GPT-2; green line: BERT; black-dashed line: human data from figure 1D as a reference). The error bars represent the estimated standard error (SE). *p<.05, **p<.01, ***p<.001.

Our behaviour results have demonstrated that the affordance of objects within the range of agent size was qualitatively different from that of objects beyond, consistent with Gibson's prediction that body size provides a metric for perceiving objects' affordance. To further examine whether only the affordance of objects within the body size is properly represented, we used fMRI to examine how the affordance of objects within and beyond the body size

were represented in regions implicated in object processing in both dorsal and ventral visual streams.

In the experiment, we measured the congruency effect of affordance when participants were instructed to evaluate whether an action was affordable by an object (Figure 3A). Four objects from the behavioral experiment were selected: two whose sizes are within the body size (i.e., bottle and football, WITHIN condition) and the other two beyond (i.e., bed and piano, BEYOND condition). Accordingly, four actions, grasp, kick, sit, and lift, were selected. During the scan, the participants were asked to decide whether a probe action was afforded (e.g., grasp-able – bottle, Congruent condition) or not (e.g., sit-able – bottle, Incongruent condition) by each of the subsequently-presented object.

We examined the congruency effect in two object-selective regions, each a representative region of the corresponding visual stream: the posterior fusiform (pFs) is involved in object recognition (e.g., Grill-Spector et al., 2000; Konkle, & Oliva, 2012; Malach et al., 1995) and objects' real-world size processing (Konkle, & Oliva, 2012; Snow et al., 2011), and the superior parietal lobule (SPL) is among the core tool network regions (e.g., Filimon et al. 2007; Matić et al., 2020). For the rest of the object-selective regions defined by the contrast of the objects by the baseline, see Supplemental Figure S3 and Supplemental Table S1. A repeated-measures ANOVA with object type (WITHIN versus BEYOND) by congruency (Congruent versus Incongruent) as within-subjects factors was performed for each ROI, separately (Figure 3C). A significant interaction between object type and congruency was observed in both ROIs (SPL: $F(1,11) = 15.47$, $p = .002$, $\eta^2 = .58$; pFs: $F(1,11) = 24.93$, $p < .001$, $\eta^2 = .69$), suggesting that these regions represented affordance differently for each object type. Pairwise comparisons further revealed a congruency effect, which the activation in the congruent condition was significantly higher than that in the incongruent condition, only for objects whose sizes were smaller than the body size (SPL: $p < .001$; pFs: $p = .021$), not for objects whose sizes larger than the body size (ps $> .41$). In addition, the main effect of object type was not significant in either ROI (ps $> .17$), suggesting that the lack of congruency effect for the objects beyond the object size was not due to the compromised engagement in viewing these objects. Further, a whole-brain analysis with the contrast of congruency for

objects beyond the object size was performed, and no region showed the congruency effect for these objects. Taken together, the affordance boundary not only separated the objects into two category based on their relative size to the body size, but also selected only the objects whose sizes were smaller than the body size for proper affordance representation in the brain.

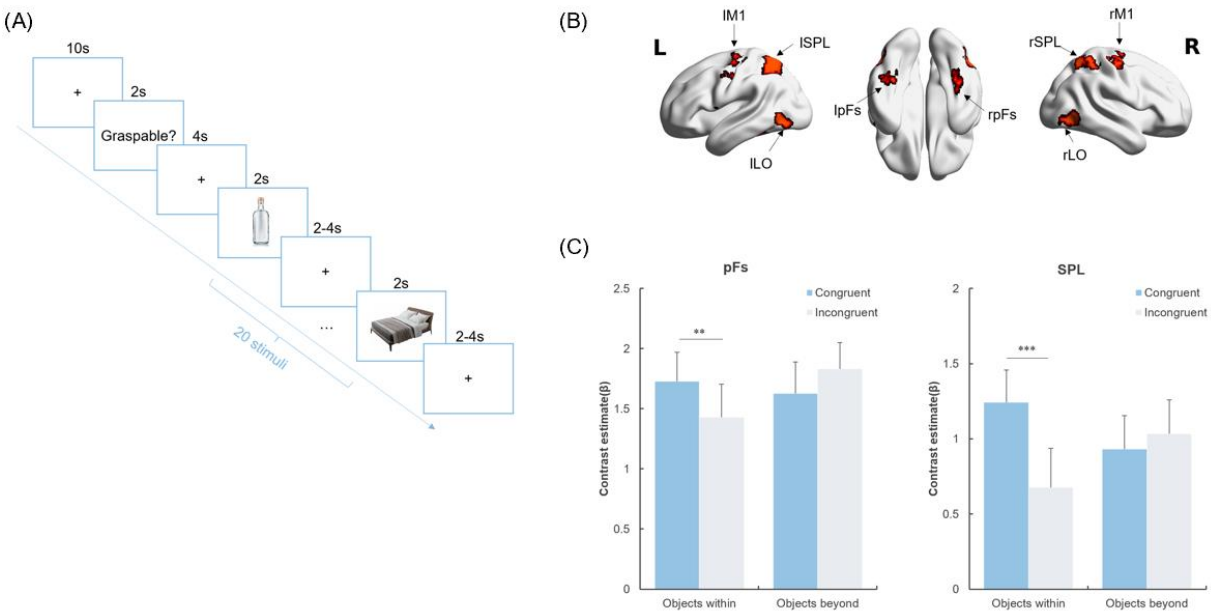

Figure 3. (A) An example block with the probe action being "graspable". Participants were asked to indicate whether each of the subsequently presented objects was graspable by pressing a corresponding button. (B) the ROIs used in this analysis; (C) the activation in each condition in pFs and SPL. The bars refer to the contrast estimates of each condition versus baseline. The stars indicate whether the contrast was significant. **p*<.05, ***p*<.01, ****p*<.001, otherwise not significant. Error bars represent the standard error (SE).

Besides the pFs and SPL, we also examined the congruency effect in the LO, which provides inputs to the pFs and SPL (Hebart et al., 2018), and the M1, which receives inputs from the dorsal stream (Vainio & Ellis, 2020) and executes object-manipulation actions (Binkofski et al., 2002). Although both the LO and M1 showed a higher response to objects than baseline, there was no congruency effect in affordance for objects within the body size

(main effect of congruency: $F(1,11) = 1.74$, $p = .214$, $\eta^2 = .13$, see supplementary Figure S4). Therefore, the representation of affordance is unlikely exclusively dictated by visual inputs and automatically engaged in motor execution, which suggests that affordance may still require some perceptual processing and not necessarily reflected in motor execution, deviating away from Gibsonian view of direct perception.

Discussion

Boundaries are a psychophysical phenomenon that highlights the discontinuity in sensation and perception in response to the continuity of physical inputs (Harnad, 1987; Young et al., 1997). This phenomenon has been demonstrated at both sensory level, such as color perception (Bornstein, & Korda, 1984) and speech-sounds perception (Liberman et al., 1957), as well as perceptual level, such as facial gender discrimination (Campanella et al., 2001). The boundaries indicate the categorical perception, which constitutes a fundamental adaptation of perception to facilitate categorizations necessary for an organism (Goldstone & Hendrickson, 2010). Our study, for the first time, unveiled a boundary in object affordance, wherein affordance similarity across the boundary was significantly lower than that within the boundary. Critically, the boundary separating objects along a size axis coincided with humans' body size, and the boundary shifted as a function of imagined body sizes when participants' body schema was directly manipulated. This implies that objects are characterized in a dimension scaled by humans' body size, and these two categories offer distinct action possibilities, which serve as a key reflection of humans' sensorimotor capacity.

Why is there a boundary in affordance perception? We propose that it is actually the boundary between (manipulable) objects and the environment. The objects whose sizes are smaller than the body size are indeed objects in the eye of an animal, and possess affordance as termed by Gibson. In contrast, the objects whose sizes are larger than the body size are not *the* "objects" in the eye of an animal ingrained to interact with the world, but a (probably uninterested) part of the "environment". To such an (Gibsonian) animal, the prerequisite of

being considered as an “object” would be to afford interaction, and an object beyond the manageable size range would not have common affordance. This speculation is in line with previous fMRI studies where large objects activate the medial part of the ventral temporal cortex (Huang et al., 2022; Magri et al., 2021), overlapped with the parahippocampus gyrus involved in representing scenes (Park et al., 2011; Troiani et al., 2014), and smaller objects activate the lateral part, such as the pFs where the congruency effect of affordance was identified. In our fMRI experiment, we found that the congruency effect of affordance is only evident for objects within the range of the body size, but not for objects beyond, suggesting that affordance is typically represented only for objects within the body size. In this sense, the body size, and the sensorimotor capacity it reflects, of the animal sets the boundary of manipulation, hence the boundary between the “things”/ (manipulable) objects and the rest (non-objects/environment).

Therefore, our study provides a new perspective on a long-lasting question in psychology, cognitive science, and even philosophy: what is an object? Studies, especially in the field of vision, define objects disembodiedly by mainly based on their physical features, such as contour (being not scrambled). Although in some studies, the manipulative definition of object adheres to layman’s idea of being a (often) detachable thing (e.g., the images used in Konkle & Oliva, 2012; or in Russakovsky et al., 2015), the line between objects, clutters and scenes remains elusive. Our finding of the affordance boundary presents a new source of object-ness: the capability of being a source of affordance under the constraint of the sensorimotor capacity of the animal. This conjecture deviates from a similar idea of Gibson that agent size sets the division between animal environment and the world (Gibson, 1979), where the animal environment is the subsection of the physical world. The environment falls into the size range of the agent, therefore is interactable and able to provides affordance to the animal. Accordingly, stimuli from the environment and the world bear ecologically distinct values to the animal, hence the dissociation between their ensured cognitive events. Our proposal is based on the same impact of the agent’s constitutional feature on its perception and interaction with the world. However, Gibson’s idea is largely a spatial division in nature, whereas we propose that the boundary set by agent size extends to the conceptual realm, and

is a source of ‘object-ness’, echoing the embodied influence on the formation of abstract concept (e.g., Barsalou, 1999; Lakoff & Johnson, 1980) of object and environment. In this way, man is indeed the measure of all things.

Although both our study and Gibsonian embodied theories of cognition emphasize the importance of the body in perceiving affordance, our study further underscores the dynamic and continuous interaction between the body schema and affordances, as the perception of affordances was sensitive to voluntary modification of body schema. Such sensitivity to voluntary control is against the directness of affordance perception that is free from cognitive mediation, arguing against early proposals in embodied cognition (e.g., Gibson, 1979, p.139; Greeno, 1994; Prindle et al., 1980). Further, the perceptual boundary can be observed in bodyless LLMs of ChatGPT, suggesting language already carries at least some information shaped by the sensorimotor information generated by the interactions during the formation and revision of concepts in language use by its original users, i.e. humans. Therefore, ChatGPT possibly learned this indirectly through a human prism, or, in our case, human metrics. This finding argues against a traditional view emphasizing the distinction between language and bodily interaction between agent and environment, as language appears to be a mental activity, derived from sensorimotor information and distilled to be verbal, conceptual, and symbolic (e.g., Gallese & Lakoff, 2005). Instead, our finding supports the grounded theories (see Barsalou, 2008 for a review) that emphasizes a bidirectional linkage between language and bodily action (see Glenberg et al., 2013; Wilson & Golonka, 2013), where the sensorimotor information from interactions may reflect in the use of languages.

Another more provocative idea is that a kind of embodiment or grounded representation automatically emerged in ChatGPT probably based on the body schema inherited from humans through language. That is, ChatGPT may learn to relate any sensorimotor-related information to a set of fixed facts (a virtual body schema), which may generate a preliminary embodied cognitive ability. In other words, ChatGPT may have derived their own algorithm to automatically use their own prism or metrics to reason the relation between bodily action and objects. Consistent with this idea, our study showed that the human-like affordance boundary became gradually obvious with the increase in the model size of the LLMs (i.e.,

greater information processing capacity). Future study is needed to test this possibility, possibly by providing sensorimotor information novel to humans, or instilling a different virtual body scheme through training corpus of language, or attaching LLMs to a real robot (Driess et al., 2023), to see if the affordance boundary observed here shifts with this altered body "metric". Taking our finding with physically embodied humans and linguistically disembodied LLMs, our findings suggest that the embodied cognition and symbolic processing of languages may be more closely and fundamentally related than we think: perception-action problems and language problems can be treated as the same kind of thing (Wilson & Golonka, 2013). In this sense, "man" is the measure of all things and all words, for humans and for AIs.

Methods

Participants

Five hundred and twenty-eight participants (311 males, aged from 16 to 73, mean age = 24.1 years) were recruited from the online survey platform (https://www.wjx.cn/), a Chinese online panel, to participate in the first behaviour task. Another one hundred participants (49 males, aged from 17 to 39 years, mean age = 23.2 years) were recruited from the same online survey platform in the second behaviour task. Each participant completed an online consent form before the experiment.

Twelve students (8 males, aged from 19 to 31 years, mean age = 23.7 years) from Tsinghua University participated in the fMRI experiment. All participants reported normal or corrected-to-normal vision. Each participant completed a pre-scan MRI safety questionnaire and a consent form before the experiment.

This study was approved by the Institutional Review Board at Beijing Normal University. All participated were all compensated financially for their time.

Stimuli

We used 27 objects from the THINGS database (Hebart et al., 2019). For each object, we selected a typical exemplar image of this daily-life object. Each image was portraited isolated against a white background, sized to 400 × 400 pixels. The objects were chosen to cover real-world size rank 2 to rank 8 as classified in Konkle and Oliva (2011), with small rank corresponds to a small real-world size (e.g., the airplane is in size rank 8 and the apple is in size rank 2). The full list of objects and the corresponding size rank was shown in Supplementary Table S2.

The stimuli for fMRI experiment consisted of images from 4 objects (bed, bottle, ball, and piano, which were a subset of those represented in the behavioural pool), with 5 exemplars for each object. The resulting 20 images (4 objects × 5 exemplars) were depicted an isolated object against a white background, all sized to 400 × 400 pixels, and presented on a screen with a resolution of 1024 × 768.

Procedure

General behavioural task

To measure the perceived affordances of objects, we created an object-action relation task, which asked the participants to map 27 objects with 14 actions. The 27 object images were pre-randomly divided into three groups (nine images each) to form a nine-box grid for display convenience. The 14 actions covered common actions that human interacts with objects and environments, referencing the actions from the kinetics human action video dataset (Kay et al., 2017).

The task comprised 42 trials (14 actions × 3 object groups) in total. In each trial, one group of object images (nine object images) and a question asking the appropriateness of applying a specific action to each object were shown (e.g., ‘Which objects are sit-able?’, see Figure 1A). Participants were asked to choose the objects that afford the specific action according to their own senses. They were told there was no right or wrong answers, as we did not presuppose a correct answer. Each object-action combination would only be presented once during the whole task. From this task, we would get the percentage that one object was

judged affording each of the 14 actions across participants. Because there has been demonstrated that there is a fundamental separation between the processing of animate and inanimate objects (e.g., Konkle & Caramazza, 2013), and the affordance of inanimate objects and that of animate are "so different" (Gibson, 1979), we only include 24 inanimate objects for the following analysis by excluding 3 animate objects (animals: bird, dog, and horse).

Behavioural task manipulating body schema

To manipulate the perceived body schema when participants engaging in the object-action related task, we asked the participants to imagine themselves to be as small as a cat, or as large as an elephant during the object-action mapping task. Each participant was randomly assigned to one body-schema condition.

The task procedure was the same to the general behaviour task, except that an illustration (Figure 1E) with an instruction "Please imagine that you have now grown smaller/larger than your real size, to roughly the same size as a cat/an elephant, as shown in the image below. Please answer the following questions based on this imagined situation." was shown before the task. The illustration was additionally presented in each trial, above the action question and the object images. By the end of the task, as a manipulation check, one additional question asked the participants to indicate their imagined body size: "What is the approximate height (cm) you imagine yourself to be during the whole task?" We only took the imagined height that were not in the average human size range (100cm - 200cm) into later data analysis.

The tasks for large language models

To test the perceived affordance of the same set of objects from the LLMs as well, we asked BERT (Bidirectional Encoder Representations from Transformers), GPT-2, and ChatGPT to complete the same object-action mapping task. Different from the human task, we had to represent nouns instead of object images to the models.

For BERT, the task was formatted as a mask filling task, in which the input was questions such as "Among airplanes, kettles, plates, umbrellas, laptops, beds, [MASK] can be

sit-able.”. We record the likelihood score that BERT provided for each of the listed objects for the masked position. For the example question, the possibility score for the word “airplane” is 0.00026 from the output list.

For GPT-2, the input questions were like “Among airplanes, kettles, plates, umbrellas, laptops, beds, the thing that can be sit-able is the [blank space]”, and we record the likelihood score GPT-2 provided for each of the listed objects in the position after the input sentence (blank space).

To mimic the sampling from human participants, we ran BERT and GPT-2 each for 20 times with different model parameters by setting different random seeds in the dropout layers, considering them as different subjects.

For ChatGPT, the task was in a direct question-and-answer format. We asked, for example, “Which objects are sit-able: ‘airplane, kettle, plate, …brick’? ” and the model answered the question by naming a subset of the object list. We then code the mentioned object afford such action as “1”, elsewise, as “0”. To get the probability for each object-action pair, we ran ChatGPT on the same task 20 times, with each newly created conversation on the OpenAI website (https://chat.openai.com/chat) considered as one subject. We then got the percentage that one object was judged affording each of the 14 actions by averaging the output from the 20 times.

Representational similarity matrix

RSM for affordance For each object, we calculated the probability that it was judged affording each of the 14 actions across participants to create a 14-dimension vector. Then, we could calculate the affordance similarity (r) between each object based on the affordance vector. A 24 × 24 symmetric matrix was then generated with the affordance similarity between object i and object j being denoted in cell (i,j). A hierarchical clustering analysis was performed and visualized with seaborn clustermap (Waskom, 2021).

RSM for objects' real-world size The similarity matrix based on the real-world size of the objects was constructed by the consistency between pairs of size ranks corresponding to each object, which was defined as (Huang et al., 2022):

$$Consistency_{ij} = 1 - \left|\frac{Rank_i - Rank_j}{8}\right|$$

where *i* and *j* are indicators to denote the object's size rank.

Affordance similarity between neighbouring size ranks

To test the relationship between object affordance and object size, we first averaged the affordance vector among objects within each size rank. Then, we calculated the correlation between the average vectors of neighbouring size ranks.

Size similarity between neighbouring size ranks

The size of each object was indexed by the real-world size in Konkle and Oliva (2011). We averaged the diagonal size within each size rank and calculated the size similarity between neighbouring size ranks, which was represented as the size change from size rank *i* to *j* relative to the average size of rank i:

$$Size\ similarity\ (i,j) = 1 - \frac{diagonal\ size_j - diagonal\ size_i}{diagonal\ size_i}$$

Sparseness of perceived affordances within and beyond agent size

To compare the patterns of action possibilities of objects within agent size and those beyond, we averaged the affordance vectors of objects in size rank 2-4 and those in size rank 5-8, respectively. The sparseness of action possibilities of each group was calculated with the following formula (Xu et al., 2021):

$$S = \frac{(\sum_{i=1,n} p_i\ /n)^2}{\sum_{i=1,n}({p_i}^2/n)}$$

where p_i is the affordable possibility of the *i*th action to such object type. The smaller the value is, the higher the selectivity is, that is to say, a smaller *S* indicates the actions afforded by the object is more dispersed.

To further validate the difference, we ran a permutation test where the *p* value for this test follows the formular adapted from Unpingco's (2016):

$$p(T > T_{obs}) = \frac{1}{N} \sum_{i=1,N} I(T_i \geq T_{obs})$$

where T_{obs} is the observed difference between the sparseness of the two types of objects, and *I* is the indicator function. Under the alpha level of 0.05, if $p < .05$, then the T_{obs} is a significant difference.

fMRI experiment design

The fMRI task design was similar to that of previous behaviour task. The fMRI scanning consisted of one high-resolution T1 anatomical run and four task runs for each participant. In each task run, participants performed four action blocks (grasp, kick, lift, and sit). The block order was counterbalanced across runs. Within each block (see Figure 3A), an introduction image was shown (2s) at the beginning to indicate the action type and followed by 20 object images (4 objects × 5 exemplars), with a jittered ISI varied between 2-4s. The object images were presented in a random order, with 2s each. The question shown on the introduction image was "Are the following presented items [grasp, kick, lift, sit]-able", corresponding to four blocks. Participants were asked to judge whether the object shown was grasp/kick/lift/sit-able or not by pressing corresponding buttons (e.g., yes: right index finger; no: left index finger). The response buttons were also counterbalanced across participants. The whole task run lasted for 464s in total, with four blocks (106s each) separated by a 10s fixation period.

With this design, we were able to measure the neural activation of objects within agent size and objects beyond agent size; further, for each object, there would be congruent trials (e.g, grasp-able – bottle: affordance = 1) and incongruent trials (e.g., sit-able – bottle: affordance = 0). We were then able to locate the brain regions representing the objects' affordance by comparing trials in which the presented objects afford the presented action

option with those do not, i.e., to locate the regions showing congruency effect (congruent - incongruent).

fMRI Data Acquisition

Imaging data were collected on a 3T Siemens Prisma MRI scanner with a 64-channel phase-arrayed head coil at the Centre for Biomedical Imaging Research in Tsinghua University. High-resolution T1-weighted images were acquired with a magnetization-prepared rapid acquisition gradient-echo (MPRAGE) sequence (TR/TE = 2530/2.27 ms, flip angle = 7°, voxel resolution = 1×1×1 mm). Functional blood-oxygen-level-dependent (BOLD) images were acquired with a T2∗-weighted gradient echo-planar sequence (TR/TE = 2000/34.0 ms, flip angle = 90°, voxel resolution = 2×2×2 mm, FOV = 200×200 mm). Earplugs were used to attenuate the scanner noise, and a foam pillow and extendable padded head clamps were used to restrain head motion. All the stimuli were projected onto a screen at the back of the scanner and were viewed from a distance of approximately 110 cm via a mirror placed on the head coil.

fMRI Data Analyses

Structural T1 and functional images were preprocessed using FSL (FMRIB's Software Library, https://fsl.fmrib.ox.ac.uk/fsl/fslwiki) v6.0.5 (ref). A standard preprocessing pipeline was applied, including skull stripping using the BET (Brain Extraction Tool, Smith, 2002), slice-timing correction, motion correction using the MCFLIRT method (Jenkinson et al., 2002), temporal high-pass temporal filtering (100s), and spatial smoothing using a Gaussian kernel of full width half magnitude (FWHM) 5mm. Each run's functional data were registered to a T1-weighted standard image (MNI152) with FLIRT.

For functional data analysis, a first-level voxel-wise general linear models (GLM) implemented in a FEAT analysis was performed on each run separately. To get neural activation maps for objects within and beyond versus baseline, the GLM included 3 regressors: objects within agent size (bottle and football), objects beyond agent size (bed and piano), and fixation period as baseline; with ISI period, response key press and introduction image included as 3 nuisance factors. The resulting first-level contrasts of parameter

estimates (COPE) were entered into the next higher-level group analyses, performed using a random-effects model (FLAME stage 1, Beckmann et al., 2003). We focused on two critical contrasts: objects within vs. fixation, and objects beyond vs. fixation, and the conjunction of these two contrasts. The resulting Z-statistic images were thresholded using the FSL's settings of $Z > 2.3$, $p = .05$ (Worsley, 2000), and corrected for multiple comparisons using adjusted cluster-wise (FWE: family-wise error) significance threshold of $p = 0.05$.

Region of interest (ROI) analysis

Eight ROIs (Figure 3B) were selected based on the overlap of the outcomes from the whole-brain conjunction map (areas activated for both objects within and beyond) and the functional atlases of these regions (pFs and LO from Zhen et al., 2015; SPL and M1 from Fan et al., 2016). For pFs and LO's atlases, the probabilistic activation maps (PAM) were thresholded at around 70%, in which map each voxel contains a percentage of participants who showed activation for seeing objects versus baseline in Zhen et al.'s study, resulting 266 voxels in lpFs, 427 voxels in rpFs, 254 voxels in lLO and 347 voxels in rLO. For SPL and M1, the probabilistic activation maps were thresholded at around 80%, resulting 661 voxels in lSPL, 455 voxels in rSPL, 378 voxels in lM1, and 449 voxels in rM1. Homologous areas within the cortical hemispheres were merged in the following ROI analysis.

Affordance congruency effect

For the affordance congruency effect of each object type, we modelled another GLM containing 5 regressors: congruent conditions for objects within/beyond, incongruent conditions for objects within/beyond, and fixation period as baseline; with ISI period, response key press and introduction image included as 3 nuisance factors. The resulting first-level COPEs were prepared for the following ROI analysis. A repeated-measures ANOVA with Object type (object within and beyond) and Congruency (congruent, incongruent) as within-subjects factors was run on the beta values (contrast estimate) extracted from their respective contrasts versus the fixation for each ROI.

To search all the possible brain regions that revealed congruency effect of objects beyond, we also ran a whole-brain analysis on the contrast between congruent condition for objects beyond vs. incongruent condition for objects beyond. The corresponding first-level COPE was entered into the group-level analyses with a random-effects model (FLAME stage 1, Beckmann et al., 2003). The resulting Z-statistic images were thresholded using the FSL's settings of Z > 2.3, p = .05 (Worsley, 2000), and corrected for multiple comparisons using adjusted cluster-wise (FWE: family-wise error) significance threshold of $p = 0.05$.

Acknowledgements

This study was funded by Natural Science Foundation of China (31600925), Beijing Municipal Science & Technology Commission, Administrative Commission of Zhongguancun Science Park (Z221100002722012), Tsinghua University Guoqiang Institute (2020GQG1016), and Beijing Academy of Artificial Intelligence (BAAI).

Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Declaration of conflicting interests

The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, *22*(4), 637-660.

Barsalou, L. W. (2008). Grounded cognition. *Annu. Rev. Psychol.*, *59*, 617-645.

Beckmann, C. F., Jenkinson, M., & Smith, S. M. (2003). General multilevel linear modeling for group analysis in FMRI. *Neuroimage*, *20*(2), 1052-1063.

Binkofski, F., Fink, G. R., Geyer, S., Buccino, G., Gruber, O., Shah, N. J., ... & Freund, H. J. (2002). Neural activity in human primary motor cortex areas 4a and 4p is modulated differentially by attention to action. *Journal of neurophysiology*, *88*(1), 514-519.

Bornstein, M. H., & Korda, N. O. (1984). Discrimination and matching within and between hues measured by reaction times: Some implications for categorical perception and levels of information processing. *Psychological research*, *46*(3), 207-222.

Campanella, S., Chrysochoos, A., & Bruyer, R. (2001). Categorical perception of facial gender information: Behavioural evidence and the face-space metaphor. *Visual cognition*, *8*(2), 237-262.

Casasanto, D. (2011). Different bodies, different minds: the body specificity of language and thought. *Current Directions in Psychological Science*, *20*(6), 378-383.

Castiello, U., Bennett, K. M. B., & Stelmach, G. E. (1993). Reach to grasp: the natural response to perturbation of object size. *Experimental brain research*, *94*(1), 163-178.

Cesari, P., & Newell, K. M. (2000). Body-scaled transitions in human grip configurations. *Journal of Experimental Psychology: Human Perception and Performance*, *26*(5), 1657.

Chemero, A. (2013). Radical embodied cognitive science. *Review of General Psychology*, *17*(2), 145-150.

Diedenhofen, B., & Musch, J. (2015). cocor: A comprehensive solution for the statistical comparison of correlations. *PloS one*, *10*(4), e0121945.

Driess, D., Xia, F., Sajjadi, M. S., Lynch, C., Chowdhery, A., Ichter, B., ... & Florence, P. (2023). PaLM-E: An Embodied Multimodal Language Model. *arXiv preprint arXiv:2303.03378*.

Fan, L., Li, H., Zhuo, J., Zhang, Y., Wang, J., Chen, L., ... & Jiang, T. (2016). The human brainnetome atlas: a new brain atlas based on connectional architecture. *Cerebral cortex*, *26*(8), 3508-3526.

Filimon, F., Nelson, J. D., Hagler, D. J., & Sereno, M. I. (2007). Human cortical representations for reaching: mirror neurons for execution, observation, and imagery. *Neuroimage*, *37*(4), 1315-1328.

Gallagher, S. (2017). Enactivist interventions: Rethinking the mind. *Oxford University Press.*

Gallese, V., & Lakoff, G. (2005). The brain's concepts: The role of the sensory-motor system in conceptual knowledge. *Cognitive neuropsychology*, *22*(3-4), 455-479.

Gibbs Jr, R. W. (2005). Embodiment and cognitive science. *Cambridge University Press*.

Gibson, J. J. (1979). The ecological approach to visual perception: classic edition.

Glenberg, A. M. (1997). Mental models, space, and embodied cognition. In *Creative thought: An investigation of conceptual structures and processes.* (pp. 495-522). American Psychological Association.

Glenberg, A. M., & Gallese, V. (2012). Action-based language: A theory of language acquisition, comprehension, and production. *cortex*, *48*(7), 905-922.

Glenberg, A. M., Witt, J. K., & Metcalfe, J. (2013). From the revolution to embodiment: 25 years of cognitive psychology. *Perspectives on psychological science*, *8*(5), 573-585.

Goldstone, R. L., & Hendrickson, A. T. (2010). Categorical perception. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(1), 69-78.

Greeno, J. G. (1994). Gibson's Affordances. *Psychological Review*, *101*(2), 336-342.

Grill-Spector, K., Kushnir, T., Hendler, T., & Malach, R. (2000). The dynamics of object-selective activation correlate with recognition performance in humans. *Nature neuroscience*, *3*(8), 837-843.

Harnad, S. (1987). Psychophysical and cognitive aspects of categorical perception: A critical overview.

Hauser, M. D., Chomsky, N., & Fitch, W. T. (2002). The faculty of language: what is it, who has it, and how did it evolve?. *science*, *298*(5598), 1569-1579.

Hebart, M. N., Dickter, A. H., Kidder, A., Kwok, W. Y., Corriveau, A., Van Wicklin, C., & Baker, C. I. (2019). THINGS: A database of 1,854 object concepts and more than 26,000 naturalistic object images. *PloS one*, *14*(10), e0223792.

Hebart, M. N., Bankson, B. B., Harel, A., Baker, C. I., & Cichy, R. M. (2018). The representational dynamics of task and object processing in humans. *Elife*, *7*, e32816.

Huang, T., Song, Y., & Liu, J. (2022). Real-world size of objects serves as an axis of object space. *Communications biology*, *5*(1), 1-12.

Jenkinson, M., Bannister, P., Brady, M., & Smith, S. (2002). Improved optimization for the robust and accurate linear registration and motion correction of brain images. *Neuroimage*, *17*(2), 825-841.

Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., ... & Zisserman, A. (2017). The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*.

Konkle, T., & Oliva, A. (2011). Canonical visual size for real-world objects. *Journal of Experimental Psychology: human perception and performance*, *37*(1), 23.

Konkle, T., & Oliva, A. (2012). A real-world size organization of object responses in occipitotemporal cortex. *Neuron*, *74*(6), 1114-1124.

Konkle, T., & Caramazza, A. (2013). Tripartite organization of the ventral stream by animacy and object size. *Journal of Neuroscience*, *33*(25), 10235-10242.

Lakoff, G., & Johnson, M. (1980). The metaphorical structure of the human conceptual system. *Cognitive science*, *4*(2), 195-208.

Liberman, A. M., Harris, K. S., Hoffman, H. S., & Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of experimental psychology*, *54*(5), 358.

Magri, C., Konkle, T., & Caramazza, A. (2021). The contribution of object size, manipulability, and stability on neural responses to inanimate objects. *NeuroImage*, *237*, 118098.

Malach, R., Reppas, J. B., Benson, R. R., Kwong, K. K., Jiang, H., Kennedy, W. A., ... & Tootell, R. B. (1995). Object-related activity revealed by functional magnetic resonance imaging in human occipital cortex. *Proceedings of the National Academy of Sciences*, *92*(18), 8135-8139.

Mark, L. S. (1987). Eyeheight-scaled information about affordances: a study of sitting and stair climbing. *Journal of experimental psychology: human perception and performance*, *13*(3), 361.

Matić, K., de Beeck, H. O., & Bracci, S. (2020). It's not all about looks: The role of object shape in parietal representations of manual tools. *Cortex*, *133*, 358-370.

Merleau-Ponty, M., & Smith, C. (1962). *Phenomenology of perception* (Vol. 2012). London: Routledge.

Newell, K. M., Scully, D. M., McDonald, P. V., & Baillargeon, R. (1989). Task constraints and infant grip configurations. *Developmental Psychobiology: The Journal of the International Society for Developmental Psychobiology*, *22*(8), 817-831.

OpenAI. ChatGPT. 2022. https://openai.com/blog/chatgpt/ (accessed Dec 21, 2022).

Park, S., Brady, T. F., Greene, M. R., & Oliva, A. (2011). Disentangling scene content from spatial boundary: complementary roles for the parahippocampal place area and lateral occipital complex in representing real-world scenes. *Journal of Neuroscience*, *31*(4), 1333-1340.

Pearson, K., & Filon, L. N. G. (1898). VII. Mathematical contributions to the theory of evolution.—IV. On the probable errors of frequency constants and on the influence of random selection on variation and correlation. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, (191), 229-311.

Prindle, S. S., Carello, C., & Turvey, M. T. (1980). Animal-environment mutuality and direct perception. *Behavioral and Brain Sciences*, *3*(3), 395-397.

Roux-Sibilon, A., Kalénine, S., Pichat, C., & Peyrin, C. (2018). Dorsal and ventral stream contribution to the paired-object affordance effect. *Neuropsychologia*, *112*, 125-134.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... & Fei-Fei, L. (2015). Imagenet large scale visual recognition challenge. *International journal of computer vision*, *115*, 211-252.

Smith, S. M. (2002). Fast robust automated brain extraction. *Human brain mapping*, *17*(3), 143-155.

Snow, J. C., Pettypiece, C. E., McAdam, T. D., McLean, A. D., Stroman, P. W., Goodale, M. A., & Culham, J. C. (2011). Bringing the real world into the fMRI scanner: Repetition effects for pictures versus real objects. *Scientific reports*, *1*(1), 1-10.

Stanfield, R. A., & Zwaan, R. A. (2001). The effect of implied orientation derived from verbal context on picture recognition. *Psychological science*, *12*(2), 153-156.

Thompson, E. (2010). *Mind in life: Biology, phenomenology, and the sciences of mind*. Harvard University Press.

Troiani, V., Stigliani, A., Smith, M. E., & Epstein, R. A. (2014). Multiple object properties drive scene-selective regions. *Cerebral cortex*, *24*(4), 883-897.

Tucker, M., & Ellis, R. (2004). Action priming by briefly presented objects. *Acta psychologica*, *116*(2), 185-203.

Unpingco, J. (2016). *Python for probability, statistics, and machine learning* (Vol. 1). Springer International Publishing.

Vainio, L., & Ellis, R. (2020). Action inhibition and affordances associated with a non-target object: An integrative review. *Neuroscience & Biobehavioral Reviews*, *112*, 487-502.

Varela, F. J., Thompson, E., & Rosch, E. (2017). *The embodied mind, revised edition: Cognitive science and human experience*. MIT press.

Warren, W. H. (1984). Perceiving affordances: visual guidance of stair climbing. *Journal of experimental psychology: Human perception and performance*, *10*(5), 683.

Warren Jr, W. H., & Whang, S. (1987). Visual guidance of walking through apertures: body-scaled information for affordances. *Journal of experimental psychology: human perception and performance*, *13*(3), 371.

Waskom, M. L. (2021). Seaborn: statistical data visualization. Journal of Open Source Software, 6(60), 3021.

Wilson, M. (2002). Six views of embodied cognition. *Psychonomic Bulletin & Review*, *9*(4), 625-636.

Wilson, A. D., & Golonka, S. (2013). Embodied cognition is not what you think it is. *Frontiers in psychology*, *4*, 58.

Worsley, K. J. (2001). Statistical analysis of activation images. *Functional MRI: An introduction to methods*, *14*(1), 251-270.

Xu, S., Zhang, Y., Zhen, Z., & Liu, J. (2021). The face module emerged in a deep convolutional neural network selectively deprived of face experience. *Frontiers in computational neuroscience*, *15*, 626259.

Young, A. W., Rowland, D., Calder, A. J., Etcoff, N. L., Seth, A., & Perrett, D. I. (1997). Facial expression megamix: Tests of dimensional and category accounts of emotion recognition. *Cognition*, *63*(3), 271-313.

Zhen, Z., Yang, Z., Huang, L., Kong, X. Z., Wang, X., Dang, X., ... & Liu, J. (2015). Quantifying interindividual variability and asymmetry of face-selective regions: a probabilistic functional atlas. *Neuroimage*, *113*, 13-25.