# Brain mechanisms of reversible symbolic reference: a potential singularity of the human brain

Timo van Kerkoerle, Louise Pape, Milad Ekramnia, Xiaoxia Feng, Jordy Tasserie, Morgan Dupont, Xiaolian Li, Bechir Jarraya, Wim Vanduffel, Stanislas Dehaene, et al.

Neuroscience

# Brain mechanisms of reversible symbolic reference: a potential singularity of the human brain

Reviewed Preprint

Published from the original preprint after peer review and assessment by eLife.

About eLife's process

**Reviewed preprint posted**
July 21, 2023 (this version)

**Sent for peer review**
March 16, 2023

**Posted to bioRxiv**
March 4, 2023

**Timo van Kerkoerle, Louise Pape, Milad Ekramnia, Xiaoxia Feng, Jordy Tasserie, Morgan Dupont, Xiaolian Li, Bechir Jarraya, Wim Vanduffel, Stanislas Dehaene, Ghislaine Dehaene-Lambertz**

Cognitive Neuroimaging Unit, CEA, INSERM, Université Paris-Saclay, NeuroSpin center, 91191 Gif/Yvette, France • Center for Brain Circuit Therapeutics Department of Neurology Brigham & Women's Hospital, Harvard Medical School, Boston, MA, USA • Department of Neurosciences, Laboratory of Neuro- and Psychophysiology, KU Leuven Medical School, Leuven 3000, Belgium • Leuven Brain Institute, KU Leuven, Leuven 3000, Belgium • Athinoula A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Charlestown, MA 02129, USA • Department of Radiology, Harvard Medical School, Boston, MA 02144, USA • Collège de France, Université Paris-Sciences-Lettres (PSL), 11 Place Marcelin Berthelot, 75005 Paris, France



## Abstract

The emergence of symbolic thinking has been proposed as a dominant cognitive criterion to distinguish humans from other primates during hominization. Although the proper definition of a symbol has been the subject of much debate, one of its simplest features is bidirectional attachment: the content is accessible from the symbol, and vice versa. Behavioral observations scattered over the past four decades suggest that this criterion might not be met in non-human primates, as they fail to generalize an association learned in one temporal order (A to B) to the reverse order (B to A). Here, we designed an implicit fMRI test to investigate the neural mechanisms of arbitrary audio-visual and visual-visual pairing in monkeys and humans and probe their spontaneous reversibility. After learning a unidirectional association, humans showed surprise signals when this learned association was violated. Crucially, this effect occurred spontaneously in both learned and reversed directions, within an extended network of high-level brain areas, including, but also going beyond the language network. In monkeys, by contrast, violations of association effects occurred solely in the learned direction and were largely confined to sensory areas. We propose that a human-specific brain network may have evolved the capacity for reversible symbolic reference.

eLife

**eLife assessment**

fMRI was used to address an **important** aspect of human cognition - the capacity for structured representations and symbolic processing - in a cross-species comparison with non-human primates (macaques); the experimental design probed implicit symbolic processing through reversal of learned stimulus pairs. The authors present **solid** evidence in humans that helps elucidate the role of brain networks in symbolic processing, however the evidence from macaques was **incomplete** (e.g., sample size constraints, potential and hard-to-quantify differences in attention allocation, motivation, and lived experience between species).

# Introduction

It is a longstanding question whether there is something unique about the cognitive abilities of humans relative to other animals (Hauser et al., 2002; Fitch et al., 2005; Iriki, 2006; Hopkins et al., 2012; Kietzmann, 2019; Penn et al., 2008; Berwick and Chomsky, 2016). Symbols are ubiquitous in many domains of human cognition, underlying not only language but mathematical, musical and social representations among many others domains (Deacon, 1998; Dehaene et al., 2022; Kabdebon and Dehaene-Lambertz, 2019; Nieder, 2009; Sablé-Meyer et al., 2021). The appearance of symbolic representations, which would develop in parallel with the expansion of prefrontal and parietal associative areas, has therefore been suggested as a crucial marker signaling hominization (Deacon, 1998; Dehaene et al., 2022; Henshilwood et al., 2002; Neubauer et al., 2018).

This proposal, however, hinges on the definition of what a symbol is. The term symbol is often used as a synonym for a sign, which is classically defined by Ferdinand de Saussure as an arbitrary binding between a “signifier” (for instance a word, a digit, but also a traffic sign, logo, etc.) and a “signified” (the meaning or content to which the signifier refers) . In that respect, however, many non-human animals, including chimpanzees, macaques, but also dogs, are able to learn hundreds of such relationships, even with arbitrary signs (Kaminski et al., 2004; Livingstone et al., 2010; Matsuzawa, 1985; Premack, 1971). Even bees can learn to associate arbitrary visual shapes to abstract representations such as visual quantities (2 or 3 elements) independently of the density, size or color of the elements in the visual display (Howard et al., 2019). More recently, it has been proposed to reserve the term “symbol” for a collection of such signs that can be syntactically manipulated according to precise compositional rules (Deacon, 1998; Dehaene et al., 2022; Nieder, 2009). The symbols then entertain relationships between each other that are parallel to the relationships between the objects, or concepts, they represent. For example, numerical symbols allow manipulations such “2+3=5” irrespective of whether it applies to apples, oranges or money. Performing the “sum” operation internally allows expectations about a specific outcome in the external world. Non-human animals may be conditioned to acquire iconic or indexical associations (i.e. signs which bear, respectively, a non-arbitrary or arbitrary relationships between the signifier and the signified) and even perhaps perform operations on the learned signs, such as addition (Livingstone et al., 2014), but their capacities for novel symbolic composition, especially of a recursive syntactic nature, appear limited, or absent (Berwick and Chomsky, 2016; Dehaene et al., 2022, 2015; Penn et al., 2008; Sablé-Meyer et al., 2021; Yang, 2013; Zhang et al., 2022).

The difference between humans and animals in terms of symbolic access remains controversial, in part because learning complex tasks require considerable training in animals, and a variety of factors such as motivation, learning rate and working memory capacity, may therefore explain an

animal's failure. This difficulty could be circumvented by testing a basic element of symbolic representations, i.e., the temporal reversibility of a learned arbitrary association. While the associations between indices and objects (typically acquired during classical conditioning) are unidirectional, as in the famous example of the whistle indicating the food, symbolic associations are bidirectional or symmetric (Deacon, 1998; Nieder, 2009). When hearing the word 'dog' for example, you can think of a dog, but when seeing a dog, you can also come up with the word 'dog'. Such reversibility is crucial for communication (the language learner must acquire both comprehension and production skills), but also for symbolic computations, which require going back-and-forth between the real world (e.g., seeing three sets of four objects), the internal symbols (e.g. to allow the internal computation "3x4=12") and back (to expect a total quantity of twelve). In the current work, we test the "reversibility hypothesis", which proposes that because of a powerful symbolic system, humans are biased to spontaneously form bidirectional associations between an object and an arbitrary sign. It implies that the referential function of the sign immediately operates in both directions (i.e., comprehension and production), allowing to retrieve the signified (meaning) from the signifier (symbol) and vice-versa.

A small number of behavioral studies, spread over four decades, report that non-human animals such as bees and pigeons, but also macaques, baboons and chimpanzees, struggle to reverse the associations that they learned in one direction (Imai et al., 2021; Kojima, 1984; Lipkens et al., 1988; Medam et al., 2016; Sidman et al., 1982; Howard et al., 2019; see Chartier and Fagot, 2022, for a review and discussion). In a recent experiment, Chartier and Fagot (2022) explored this question in 20 free-behaving baboons. After having learned to pair visual shapes (two pairs A-B) above 80% success, their performance dropped considerably when the order of presentation was subsequently reversed (B-A; 54% correct, chance = 50%), although their relearning performance was only slightly but significantly better when the reversed pairs were congruent (B1-A1; B2-A2) rather than incongruent (B1-A2; B2-A1). Even for the famous case of chimpanzee *AI*, who learned Arabic numerals and other arbitrary tokens for colors and objects (Matsuzawa, 2009, 1985), it turns out that her capacity to associate signs and their meanings was based on an explicit and sequential training in both directions, at least initially (Kojima, 1984). In sharp contrast, humans as young as 8 months, even when tested under the same conditions as monkeys or baboons (Sidman et al., 1982), show behavioral evidence of immediate spontaneous reversal of learned associations (Imai et al., 2021; Ogawa et al., 2010; Sidman et al., 1982).

Still, behavioral tests depend on an explicit report which could hide an implicit understanding of symbolic representations. This confound can be alleviated by directly recording the brain responses, providing a more direct comparison between species. Here, we propose a simple brain-imaging test of reversible associations. First, the participant receives evidence of several stimulus pairings between an object (O) and an arbitrary sign or label (L) in a fixed 'canonical order', e.g., from $O_1$ to $L_1$ and from $O_2$ to $L_2$. Knowledge of these learned (i.e., congruent) associations is then tested using a classic violation-of-expectation paradigm, by evaluating the brain's surprise response or "prediction error" when, say, $O_1$ is followed by $L_2$. This response can then also be evaluated in the converse direction, by switching the order of presentation of the two items within a pair. The crucial question is whether the brain shows a surprise response to an incongruent pairing presented in reversed order (e.g., $L_1$ followed by $O_2$), relative to the corresponding congruent pairing ($L_1$ followed by $O_1$). The reversibility hypothesis predicts that if symbolic associations are formed, pairs presented in canonical and reversed order should be similarly processed, and so a similar surprise response to incongruent pairings should be found in both cases.

A recent study from our lab used EEG to apply this approach to 4-5 month-old human infants (Kabdebon and Dehaene-Lambertz, 2019). The infants were habituated to pairs of stimuli in which a specific picture (a lion or a fish) was associated with tri-syllabic none words, depending on a rule concerning syllable- repetition in the word (e.g. xxY words such as *babagu*, *didito*, etc..

were followed by the fish picture whereas xYx words such as *lotilo*, *fudafu*, etc.. were followed by the lion picture). Violation-of-expectations responses were recorded in both canonical and reverse order, suggesting that preverbal human infants, already have the ability to reversibly attach a symbol to an abstract rule. In human adults, an fMRI study with a more complex design using explicit reports on associations between abstract patterns also showed brain signatures suggestive of spontaneous reversal of learned associations (Ogawa et al., 2010). The network of brain areas overlapped with the multiple-demand system that is ubiquitously observed in high- level cognitive tasks (Duncan, 2010; Fedorenko et al., 2013), including bilateral inferior and middle frontal gyrus (IFG and MFG), anterior insula (AI), intraparietal sulcus (IPS), and dorsal anterior cingulate cortex (dACC). In contrast, a human fMRI study investigating association learning between two natural visual objects found that violation effects in the learned direction were restricted to low level visual areas (Richter et al., 2018). Similarly, in macaque monkeys violation effects in the learned direction have been found selectively in visual areas, using fMRI as well as single-neuron recordings (Kaposvari et al., 2018; Meyer et al., 2014; Meyer and Olson, 2011; Vergnieux and Vogels, 2020). One of these studies (Meyer and Olson, 2011) also tested, in a small subset of 17 neurons, whether the learned associations spontaneously reversed, and showed no such reversal. From these studies, it is difficult to draw a conclusion about a potential difference between species, due to important differences in recording techniques and task design.

Here, we directly compared the ability to spontaneously reverse learned associations in humans and macaque monkeys using identical training, stimuli and whole-brain fMRI measures. Our goals were to (1) probe the reversibility hypothesis in an elementary passive paradigm in both species; (2) to shed light on the brain mechanisms of symbolic associations in humans. Indeed, two alternative hypotheses may be formulated. First, given that symbolic learning is a defining feature of language, reversible violation-of- expectation effects might be restricted to the left-hemispheric temporal and inferior frontal language areas. Alternatively, since symbolic learning is manifest in many domains outside of language, for instance in mathematics or music, each attached to a dissociable fronto-posterior brain network (Amalric and Dehaene, 2016; Chen et al., 2021; Dehaene et al., 2022; Fedorenko et al., 2011; Nieder, 2019; Norman-Haignere et al., 2015), reversibility could be expected to arise from a broad and bilateral network of human brain areas, including dorsal intraparietal and middle frontal nodes. We thus tested audio-visual and visual-visual symbolic pairing in two successive experiments.

# Results

## Summary of the experimental design

In the first experiment, we examined the learning and reversibility of auditory-visual pairs, i.e., between a visual object and an auditory label. Over the course of three days, we habituated humans (n=31) and macaque monkeys (n=2) with 4 pairs of visual objects and speech sounds (**Figure 1A**; **Supplementary Figure 1**). Two of the pairs were presented in the auditory-to-visual direction and two in the visual-to-auditory direction, ensuring that all subjects had experience with both orders and would not be surprised by their temporal reversal *per se* (see discussion of the utility of this point in Medam et al, 2016). After three consecutive days of habituation with 100% of congruent canonical trials (24 training trials in total per pair, presented outside the scanner), subjects were tested for learning using 3T fMRI, during which they were passively exposed to pairs that respected or violated the learned pairings (**Figure 1B**). To sustain the memory for learned pairs, the design still included 70% of congruent canonical trials (identical to the trials presented during habituation). In addition, there were 10% of incongruent canonical trials, in which the temporal order was maintained but the pairings between auditory and visual stimuli were violated. Enhanced brain responses to such incongruent pairs would indicate surprise and therefore prove that the associations had been learned. Note that all auditory and visual stimuli themselves were familiar: only their pairing was unusual. The design also included

10% of reversed congruent and 10% of reversed incongruent trials, in which the habitual (i.e. canonical) order of presentation of the pairs was reversed (**Figure 1A**). Observing an incongruity effect on such reversed trials would indicate that subjects spontaneously reversed the pairings and were surprised when they were violated. Note that the frequency of the two types of reversed trials was equal, and thus did not afford any additional learning of the reversed pairs (unlike Chartier and Fagot, 2022).

## Experiment 1| audio-visual stimulus pairs

We first mapped the cortical regions that were activated by visual and auditory stimuli, modelling the two stimuli within each pair with separate regressors (**Figure 1B, C**). Visually evoked activations propagated all the way to the prefrontal cortex (PFC) in monkeys while they remained restricted to lower cortical areas in humans, in line with previous studies (Denys et al., 2004b; Mantini et al., 2013). In contrast, the response was relatively weak in the auditory cortex of monkeys, also in line with previous studies (Erb et al., 2019; Petkov et al., 2009; Uhrig et al., 2014). This is expected as the size of the auditory cortex in monkeys is small, relative to their visual cortex (Felleman and Van Essen, 1991), as well as relative to the size of the auditory cortex in humans (Woods et al., 2010). Even though the onset of the two stimuli within a pair were just 800ms apart, the fast acquisition allowed us to separate the timing of the activation of the visual and auditory pathways in both humans and monkeys (**Figure 1 D, E**). In visual cortex, the response evoked by the pair arose earlier when the first stimulus of the pair was visual compared to when it was auditory, and the other way around for the auditory cortex.

We next investigated whether the subjects had learned the associations, whether the brain responses showed signatures of generalization to the reversed direction, and which brain areas were involved. If participants had learned the associations, incongruent trials should evoke a surprise response relative to congruent trials, when presented in the same order as the training pairs (canonical trials). Crucially, if they spontaneously reversed the associations, a similar incongruity effect should also be seen on reversed trials. According to the reversibility hypothesis, humans should show a spontaneous reversal, while monkeys should not. Only for monkey, we should therefore find an interaction effect between incongruity and canonicity, indicating a significant difference between the congruity effect in the learned direction compared to the congruity effect in the reserved direction.

Indeed, in humans, a vast network was activated by incongruity on both canonical and reversed trials (voxel $p<0.001$, cluster $p<0.05$ corrected, n=31 participants) (**Figure 2A**, **Table 1**). This network included a set of high-level brain regions previously described as the multiple demand system (Duncan, 2010; Fedorenko et al., 2013), including bilateral IFG, MFG, AI, IPS, and dACC. It also included the language network (Pallier et al., 2011), with the left superior temporal sulcus (STS), and the left IFG. However, in our case the activation was bilateral, thereby supporting the model that the language network is part of a larger symbolic network (Dehaene et al., 2022). Furthermore, we also found activations in the precuneus, similar to the network that has been found for top-down attention to memorized visual stimuli (Sestieri et al., 2010), which also included bilateral STS and IPS. Notably, we did not find any congruity effects in visually activated regions (compare to **Figure 1B**), in contrast to a previous human fMRI study (Richter et al., 2018). **Figure 2B** shows the hemodynamic response within the different clusters and the different conditions. In all analyses, since there were a majority of canonical congruent trials, sensitivity was higher in the canonical direction, and thus the size of the significant clusters was larger on canonical than on reversed trials. However, no significant cluster exhibited any interaction between congruity and canonicity, indicating that there was no statistical difference between the effect of congruity for the habituated and the reversed direction. Thus, the human brain fully and spontaneously reverses the auditory-visual associations that it learns.

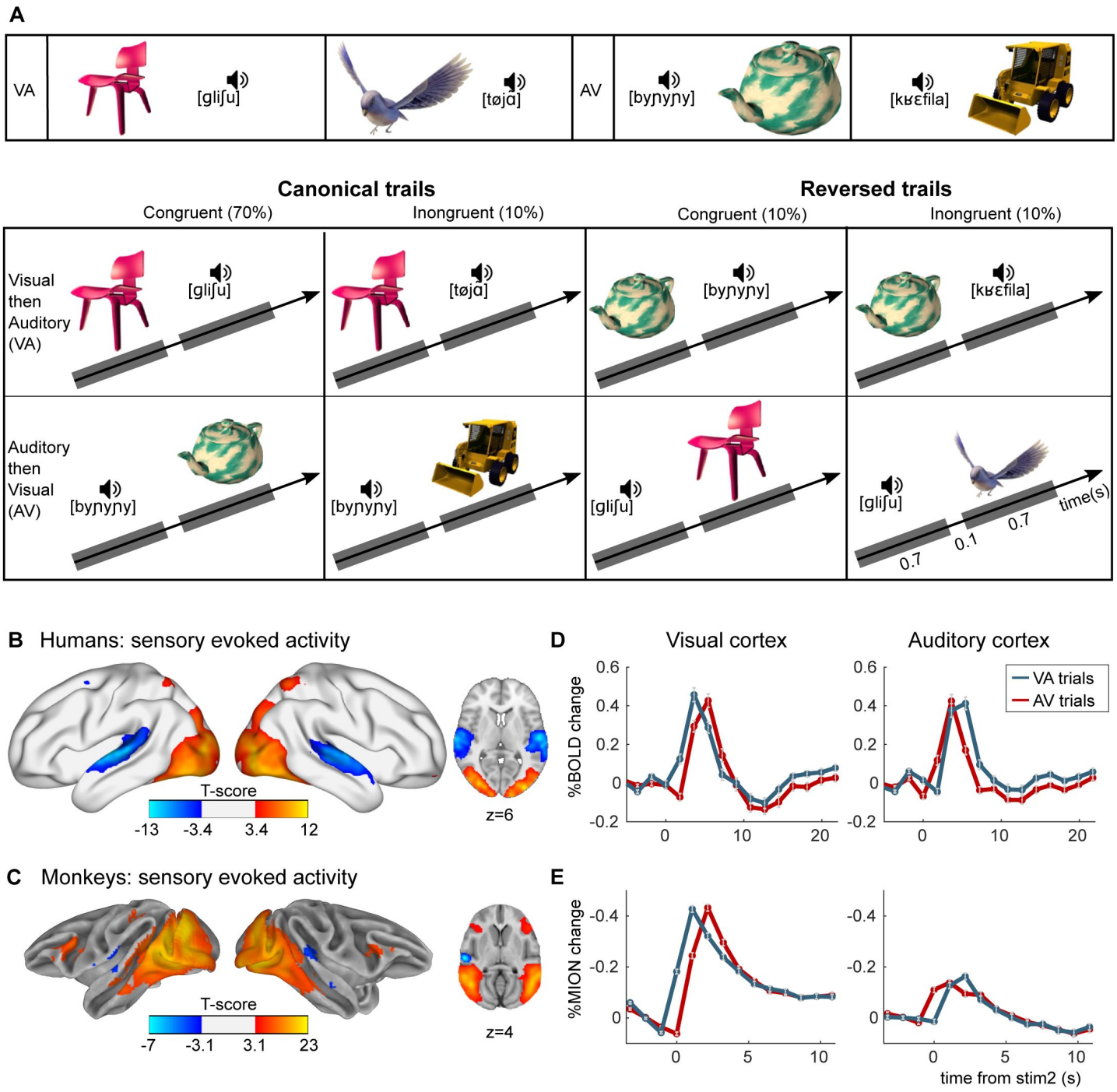



**Figure 1.**

**Experimental paradigm for auditory-visual label learning.**

**A**) Subjects were exposed to four different visual-auditory pairs during three days (6 repetitions of each pair, 3 minute video). Two pairs were always presented in the 'visual-then-auditory' order (object to label), and two in the 'auditory-then-visual' (label to object) order. During the test phase, this canonical order was kept on 80% of trials, including 10% of incongruent pairs to test memory of the learned pairs, and was reversed on 20% of the trials. On reversed trials, half the pairs were congruent and half were incongruent (each 10% of total trials), thus testing reversibility of the pairings without affording additional learning. **B**,**C**) Activation in sensory cortices. Although each trial comprises auditory and visual stimuli, these could be separated by the temporal offsets. Images show significantly activated regions in the contrasts image > sound (red-yellow) and sound > image (blue-light blue), averaged across all subjects and runs for humans (B) and monkeys (C). **D**,**E**) Average finite-impulse-response (FIR) estimate of the deconvolved hemodynamic responses for humans (D) and monkeys (E) within clusters shown in B and C respectively, separately for visual-audio (VA) and audio-visual (AV) trials. Sign flipped on y-axis for monkey responses.

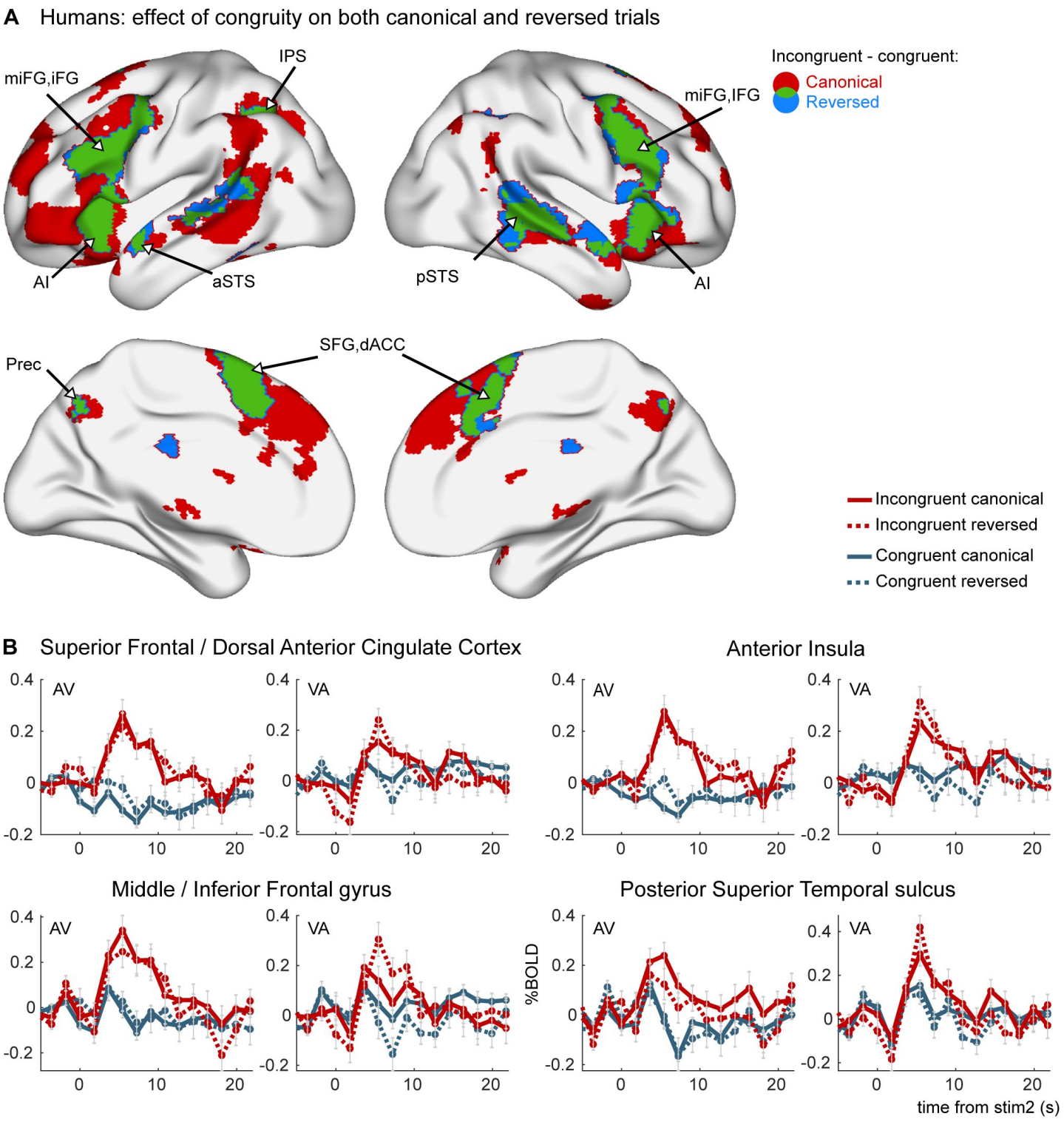



**Figure 2.**

**Congruity effects in the auditory-visual task in humans (experiment 1).**

**A**) areas activated by incongruent trials more than by congruent trial in canonical trials (red), reversed trials (blue), and their overlap (green). Brain maps are thresholded at $p_{voxel} < 0.001$ & $p_{cluster} < 0.05$ corrected for multiple comparisons across the brain volume. No interaction effect was observed between congruity and canonicity. **B**) Average FIR estimate of the deconvolved hemodynamic responses within significant clusters in the left hemisphere, separately for VA and AV trials.

| Region | MNI coordinates | Congruity effect (t-values) | | |
|---|---|---|---|---|
| | | Main | Canonical trials | Reversed trials |
| L sup frontal | -26 56 24 | 4.40 | 4.41 | 2.10 |
| L precentral | -36 6 32 | 5.75 | 3.57 | 7.50 |
| L triangularis | -48 16 2 | 7.65 | 5.45 | 6.08 |
| L insula | -40 22 0 | 7.76 | 5.84 | 6.27 |
| L temporal pole | -60 2 -10 | 6.56 | 3.95 | 5.71 |
| L ant STS | -62 -24 0 | 5.71 | 4.28 | 4.09 |
| L post STS | -54 -34 4 | 4.82 | 2.78 | 5.09 |
| L precuneus | -6 -68 40 | 4.68 | 4.72 | 3.39 |
| L inf parietal | -28 -58 42 | 5.85 | 3.97 | 4.56 |
| L caudate | -10 2 14 | 5.22 | 5.15 | 3.03 |
| L cerebellum | -6 -82 -34 | 5.59 | 3.98 | 3.27 |
| R mid frontal | 54 26 32 | 7.79 | 5.34 | 5.86 |
| R opercularis | 50 20 32 | 7.32 | 5.44 | 6.74 |
| R insula | 40 22 0 | 5.83 | 4.93 | 5.11 |
| R temporal pole | 60 4 -14 | 6.89 | 5.52 | 4.49 |
| R post STS | 48 -32 0 | 7.48 | 5.96 | 5.47 |
| R precuneus | 4 -62 40 | 6.36 | 5.16 | 2.88 |
| R inf parietal | 34 -64 44 | 5.14 | 3.57 | 4.49 |
| R caudate | 10 2 14 | 4.21 | 4.35 | 2.67 |
| R cerebellum | 10 -76 -24 | 5.06 | 5.06 | 2.02 |

R:Right; L: Left; STS: Superior temporal Sulcus

**Table 1**

Congruity effect in Experiment 1 in humans (n=31).

We next asked whether monkeys (n=3) also learned the associations and did so in both directions. The canonical congruity effect, indexing learning, was not significant when analyzing only the first imaging session after the 3 days of training. Thus, monkeys were further trained during two weeks (with in total ~960 training trials per pair) and tested during 4 consecutive days. The same training and testing pattern was used for 5 stimuli sets (**Supplementary Figure 1**). After this extended training, we found consistent effects in both monkeys, with clusters in early visual areas (V1, V2, V4), and auditory association areas in the left temporo-parieto-occipital cortex (TPO) (AV and VA trials combined, p>0.001, cluster p<0.05, n=2) (**Figure 3**, **Table 2**). Crucially, however, this effect was confined to the canonical direction, with no significant clusters in the reversed direction at the whole-brain level, in accordance with the reversibility hypothesis. We specifically tested the difference between the congruity effect in the learned and the reversed direction by calculating the interaction effect between congruity and canonicity, which showed an activation pattern that was similar to the canonical congruity effect, which reached significance in areas V2 and V4. **Figure 3C** shows the corresponding hemodynamic signals, with an enhanced response to incongruent pairs in the canonical direction (continuous red curve) but not in the reversed direction (dashed red curve). The results thus indicated that monkey cortex could acquire audio-visual pairings, as also shown by prior visual-visual experiments (Meyer and Olson, 2011; Vergnieux and Vogels, 2020), but with two major differences with humans: the congruity effects did not involve a broad network of high-level cortical areas but remained restricted to early sensory areas, and the learned associations did not reverse.

## Experiment 2 | visual-visual stimulus pairs

The non-reversal in monkeys in the above audio-visual experiment could be due to a number of methodological choices. First, although the visual stimuli were optimized for monkeys, as 3 out of 5 stimulus sets were pictures of familiar toys, the auditory stimuli (pseudowords) might have been suboptimal for them (although note that monkeys in our lab have extensive experience with human speech). It might be argued that this choice made their discrimination difficult (although note that the canonical congruity effect is evidence of discrimination). Indeed, the auditory cortex is relatively small in monkeys compared to humans (Woods et al., 2010), and there is evidence that auditory memory capacity is reduced in monkeys compared to humans (Scott and Mishkin, 2016). Second, the instructions differed: while we asked human subjects to fixate a dot at the center of the screen and to pay attention to the stimuli, monkeys were simply rewarded for fixation.

To address those concerns, we replicated the experiment with reward-dependent visual-visual associations in 3 macaque monkeys (**Figure 4**; **Supplementary Figure 2A**). First, we replaced the spoken auditory stimuli with abstract black-and-white shapes similar to the lexigrams used to train chimpanzees to communicate with humans (Matsuzawa, 1985) (**Supplementary Figure 2B**). Second, to enhance attention for the monkeys, we introduced a reward association paradigm that made the stimuli behaviorally relevant for them (Wikman et al., 2019). Within each presentation direction, one of the two pictures of objects was associated with a high reward, and one with a low reward (**Supplementary Figure 2A**). Monkeys were still rewarded for fixation, but object identity predicted the size of the reward during the delay period following the presentation of the stimuli (two objects predicted a high reward, and two predicted a low reward). To calculate congruity effects, the two pairs within each direction were always averaged, making the reward association an orthogonal element in the design.

Using this design, we obtained significant canonical congruity effects in monkeys on the first imaging day after the initial training (24 trials per pair), indicating that the animals had learned the associations (**Figure 4B**, **Table 3**). The effect was again found in visual areas (V1, V2 & V4), also spreading to the prefrontal cortex (45B, 46v), very similar to the visually activated areas (compare to **Figure 1C**). In addition, small clusters were also found in area 6 and in STS. . Crucially, the congruity effect remained restricted to the learned direction, as no area showed a

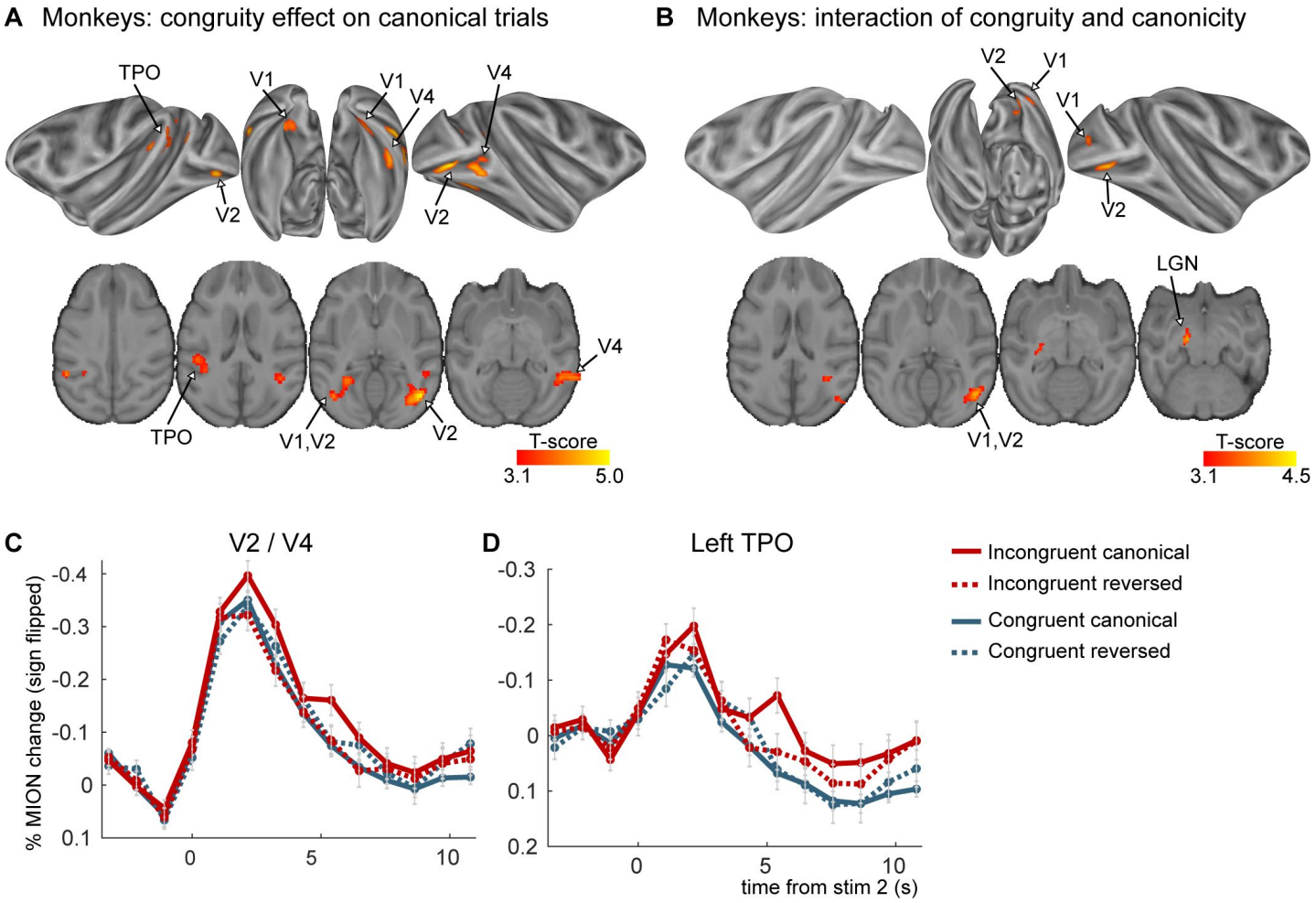



**Figure 3.**

**Congruity effects in the auditory-visual task in monkeys (experiment 1).**

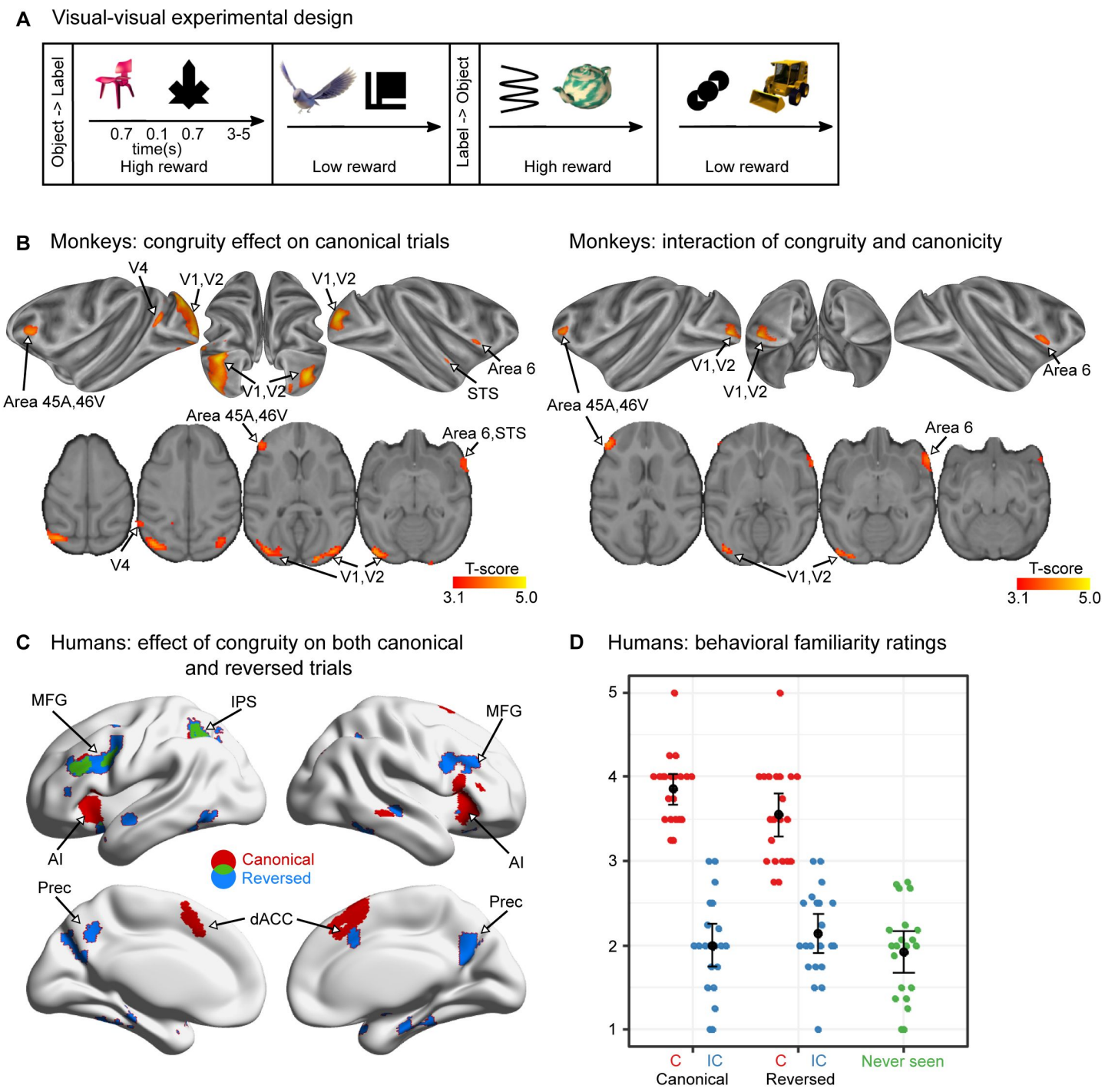
**A**) significant clusters from the incongruent-congruent canonical contrast. No significant clusters were found for the reversed direction. **B**) significant clusters from the interaction between congruity and canonicity. ($p_{voxel}$<0.001 & $p_{cluster}$<0.05 for both maps) **C**,**D**) Average FIR estimate of the deconvolved MION responses within the clusters from the incongruent-congruent canonical contrast, averaged over VA and AV trials. All clusters in early visual areas were taken together to create figure C. Average of 2 animals.

| region | MNI coordinates | Congruity effect canonical trials (t-values) | Interaction effect : congruity x canonicity (t-values) |
|---|---|---|---|
| R V2, V4 | 17 -29 4 | 5.04 | 4.56 |
| L V2 | -18 -30 2 | 4.6 | <1 |
| R V4 | 21 -22 0 | 4.23 | <1 |
| L TPO | -20 -21 11 | 4.13 | <1 |
| L LGN | -8 -8 -5 | <1 | 3.98 |

R = Right; L = Left; TPO = temporo-parieto-occipital cortex; LGN = lateral geniculate nucleus

**Table 2**

Congruity effect in Experiment 1 in monkeys (n=2)




## Figure 4.

### Visual-visual label learning in humans and monkeys (experiment 2).

**A, Experiment paradigm.** Subjects were habituated to 4 different visual-visual pairs during three days. Two pairs were in the 'object-then-label' order and two pairs in the 'label-then-object' order. For the monkeys, one object in each direction was associated with a high reward while the other one was associated with a low reward, making reward size orthogonal to congruity and canonicity (See **Supplementary Figure 2** for details). **B, monkey fMRI results.** Significant clusters ($p_{voxel}$<0.001 & cluster volume >50) from the incongruent- congruent canonical contrast (left) and the interaction between congruity and canonicity (right). **C, human fMRI results.** Areas more activated by incongruent trials more than by congruent trial in canonical trials (red), reversed trials (blue), and their overlap (green) (right) ($p_{voxel}$<0.005 & cluster volume >50). No red voxels are visible because all of them figure in the overlap (green). **D, Human behavioral results**. After learning, human adults rated the familiarity of different types of pairs (including a fifth category of novel, never seen pairings). Each dot represents the mean response of one subject in each condition. Although the reversed congruent trials constituted only 10% of the trials, they were considered almost as familiar as the canonical congruent pairs.

significant reversed congruity effect, again in accordance with the reversibility hypothesis. The interaction between congruity and canonicity indicated that there was a significant difference between the canonical and the reversed direction in a similar set of regions (V1, V2, area 45A, 46v and 6). The greater involvement of frontal cortex in the congruity effect in this paradigm fits with previous reports on the impact of reward association on long-term memory for visual stimuli in macaque monkeys (Ghazizadeh et al., 2018) . To further investigate this, we split high versus low rewarded pairs and found that congruity effect was present only for high-reward conditions, with a significant interaction of congruity and reward in area 45 and caudate nucleus (**Supplementary Figure 3**). Overall, these results indicate that, even when stimuli were optimized and made relevant for monkeys, leading to enhanced activations and an activation of prefrontal cortex to violations of expectations, the learned associations did not reverse in monkeys.

We also ran this visual-visual paradigm in human participants (n=24) with the goal to clarify the role of language in the reversibility process. Humans again gave evidence of reversed association, although weaker than with spoken words (**Figure 4C**, **Table 4**). At the normal threshold (voxel p<0.001, cluster p<0.05 corrected), the main effect of congruity was significant in a network very similar to experiment 1, including bilateral middle frontal gyrus (MFG), left intraparietal sulcus (IPS), bilateral anterior insula, dorsal anterior cingulate cortex (dACC), with an additional focus in left inferior temporal gyrus (**Figure 4C**, **Table 4**). The involvement of the language network was limited. In particular a main effect of congruity in the STS was absent, in agreement with the shift to visual symbols. Still, bilateral middle frontal gyri, STS and the precuneus were again activated by the incongruent minus congruent contrast on reversed trials (voxel p<0.001, cluster p<0.05 corrected), thereby extending beyond the multiple-demand system (Duncan, 2010; Fedorenko et al., 2013). While sensory activated regions were again absent, in contrast to a previous study on congruity effects in humans when using associations between two visual objects (Richter et al., 2018). And crucially, no interaction effect was again found between congruity and canonicity, neither at the classical threshold (p<0.001) nor at a lower threshold (p<0.01). Those results indicate that humans can also encode pairs of visual stimuli in a symmetrical, reversible fashion, involving a network of high-level cortical areas, unlike monkeys.

Further evidence was obtained from a behavioral test, performed after imaging, where we collected familiarity ratings for each stimulus pair (see Methods, **Figure 4**). Although participants reported a higher familiarity with congruent canonical pairs (which were presented on 70% of trials) than with congruent reversed pairs (which were presented on 10% of trials, t(20)=2.8, p=0.01), both pairs were rated as much more familiar than their corresponding incongruent pairs (although they were also presented 10% of time), and than never-seen pairs (all t(20) >7, p<0.0001, bilateral paired t-test). This familiarity task thus confirms that humans spontaneously reverse associations and experience a memory illusion of seeing the reversed pairs.

## Joint analysis of audio-visual and visual-visual stimulus pairs

In order to better characterize the human reversible symbol learning network and its dependence on modality, we reanalyzed both human experiments together (n=55) (**Supplementary Figure 4**). There was, unsurprisingly, a main effect of experiment with greater activation in a bilateral auditory and linguistic network in the AV experiment, and in the occipital, occipito-temporal and occipito-parietal visual pathways in the VV experiment. A main effect of congruity was observed and was again significant in both directions, canonical and reversed, in bilateral regions: insula, MFG, precentral, IPS, precuneus, ACC and STS. Crucially, there was still no region sensitive to the congruity X canonicity interaction, indicating that the learned associations were fully reversible. Finally, a single region, the left posterior STS, showed a significantly different congruity effect in the two experiments, as it was slightly larger in the AV relative to VV paradigm ([-60 -40 8], z=4.51; 183 vox, pcor=0.049), compatible with a specific role in learning of new spoken lexical items. The results therefore suggest that a broad and bilateral network, encompassing language areas but extending beyond them into dorsal parietal and prefrontal cortices, responded to violations of reversible symbolic association regardless of modality.

| region | MNI coordinates | Congruity effect canonical trials (t-values) | Interaction effect : congruity x canonicity (t-values) |
|---|---|---|---|
| L V1, V2 | -17 -36 1 | 5.18 | <1 |
| R V1, V2 | 15 -35 7 | 4.76 | <1 |
| L V4 | -23 -23 8 | 3.92 | <1 |
| L area 45A, 46v | -17 14 6 | 3.89 | <1 |
| MNIR STS | 22 6 -3 | 3.65 | <1 |
| L TPO | -8 -17 13 | 3.45 | <1 |
| L V1/V2 | -18 -35 1 | <1 | 4.83 |
| L area 45A, 46v | -17 15 6 | <1 | 4.04 |
| R area 6 | 23 7 -1 | <1 | 3.91 |

R = Right; L = Left; STS = superior temporal sulcus; TPO = temporo-parieto-occipital cortex

**Table 3**

Congruity effect in Experiment 2 in monkeys (n=3)

| Region | MNI coordinates | Congruity effect (t-values) Main | Canonical trials | Reversed trials |
|---|---|---|---|---|
| L triangularis | -44 30 24 | 5.34 | 3.64 | 3.91 |
| | -34 26 0 | 4.43 | 4.36 | 1.91 |
| L ant cingulaire | -8 18 42 | 4.52 | 3.25 | 3.13 |
| L suppl motor area | 2 20 52 | 3.79 | 3.95 | 1.40 |
| L precentral | -48 4 40 | 4.82 | 2.56 | 4.26 |
| L inf parietal | -30 -50 44 | 5.09 | 3.90 | 3.30 |
| L mid occipital | -28 -70 32 | 5.05 | 2.89 | 2.79 |
| L visual word form area | -50 -60 -12 | 4.43 | 2.62 | 3.64 |
| R sup frontal | 56 24 36 | 4.93 | 3.41 | 3.57 |
| R orbito frontal | 26 26 -16 | 5.05 | 1.92 | 5.22 |
| R operculum | 50 16 -2 | 3.58 | 2.96 | 2.11 |
| | 48 10 28 | 4.74 | 2.20 | 4.39 |

R: Right; L: Left;

**Table 4**

Congruity effect in Experiment 2 in humans (n=23)

To interrogate more finely the role of language-related and non-related areas, we turned to a sensitive subject-specific region-of-interest (ROI) analysis. We selected ROIs which are considered as the main hubs of language (Pallier et al., 2011), mathematics (Amalric & Dehaene, 2016) and reading networks. Within these ROIs, we used a separate localizer (Pinel et al., 2007) to recover the subject-specific coordinates of the 10% best voxels involved in amodal sentence processing (within language ROIs), in mental arithmetic (within mathematical ROIs), and in sentence reading relative to listening (within the visual word form area, VWFA). We added this region as it is activated by written words, visual symbols *par excellence*. We then performed ANOVAs on the betas of the main experiment averaged over these voxels.

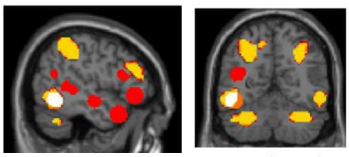
A main congruity effect was observed in all ROIs (**Table 5**). There was also a main effect of experiment in all language ROIs, VWFA and right IT, due on the one hand to larger activations in the AV than VV experiment in frontal and superior temporal ROIs, and on the other hand to the converse trend in the VWFA and IT ROIs. A significant congruity x experiment interaction was seen only in the pSTS and IFG triangularis, because these ROIs showed a large congruity effect in the AV experiment, but no effect in the VV experiment – thus further confirming that these areas contribute specifically to the acquisition of linguistic symbols, while all other areas were engaged regardless of modality. Importantly, in all these analyses, no significant interaction canonicity X congruity nor experiment x canonicity X congruity were observed, confirming the whole brain analyses (**Supplementary Figure 4** and **Table 5**).

Finally, in experiment 2 in which participants rated the familiarity of the pairs, we computed a within- subject behavioral index of reversibility as the difference in familiarity rating between incongruent and congruent reversed pairs. Across subjects, this index was correlated with the fMRI congruity effect (difference between incongruent and congruent trials in the ROI) on canonical trials (r=0.49, p=0.028) and especially on reversed trials (r=0.64, p=0.002) in the left dorsal part of area 44. In the right cerebellum, a similar correlation was observed but only for the reversed trials (r=0.57, p=0.008). No significant correlation was observed in other ROIs.

# Discussion

Using fMRI in human and non-human primates, we studied the learning of a sequential association between either a spoken label and an object (Exp. 1), or a visual label and an object (Exp. 2). In humans, we observed no difference in brain activation between the learned and the temporally reversed associations: in both directions, violations of the learned association activated a large set of bilateral regions (insula, prefrontal, intraparietal, cingulate cortex,) that extended beyond the language processing network. Thus, humans generalized the learned pairings across a reversal of temporal order (**Figure 5**). In contrast, non-human primates showed evidence of remembering the pairs only in the learned direction and did not show any signature of spontaneous reversal. Monkey responses to incongruent pairings were entirely confined to the learned canonical order and occurred primarily within sensory areas, with propagation to frontal cortex only for rewarded stimuli, yet still only in the forward direction (**Figure 5**).

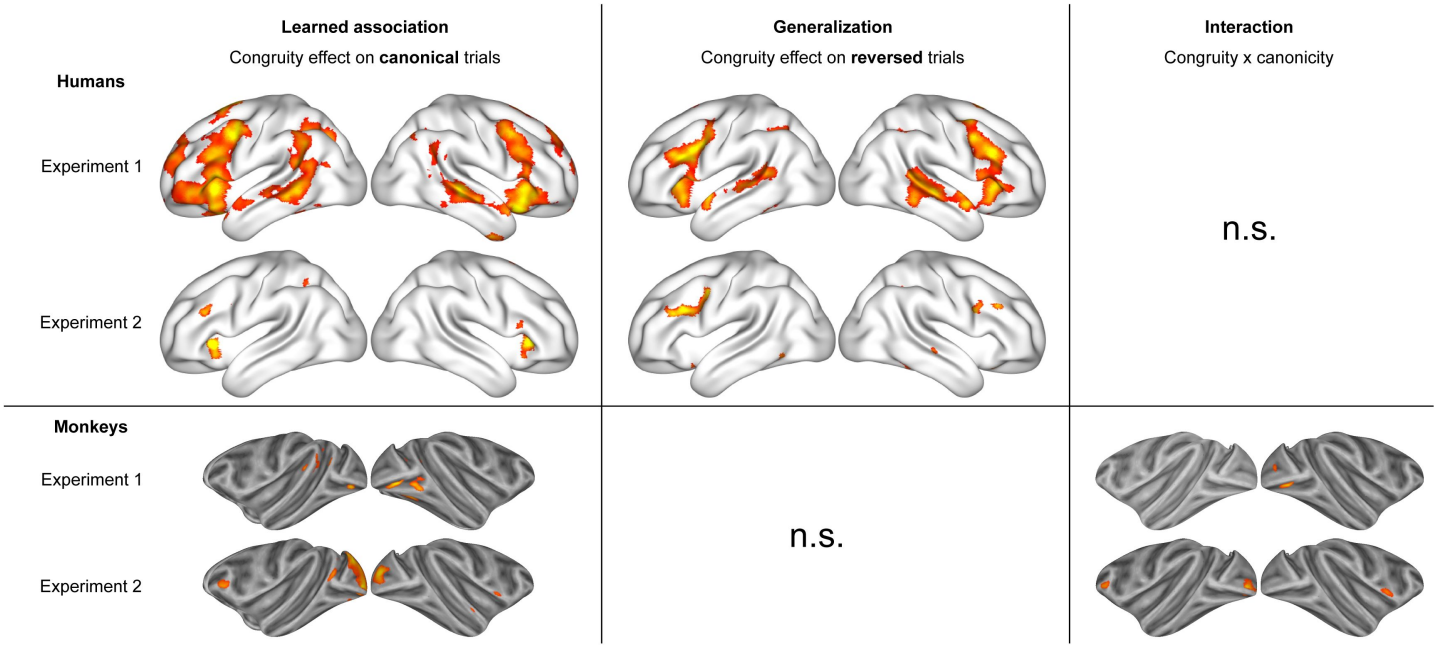
Several studies previously found behavioral evidence for a uniquely human ability to spontaneously reverse a learned association (Imai et al., 2021; Kojima, 1984; Lipkens et al., 1988; Medam et al., 2016; Sidman et al., 1982), and such reversibility was therefore proposed as a defining feature of symbol representation reference (Deacon, 1998; Kabdebon and Dehaene-Lambertz, 2019; Nieder, 2009). Here, we went one step further by testing this hypothesis at the brain level. Indeed, a limit of previous behavioral studies is that animals could have understood the reversibility of a symbolic relationship, but failed to express it behaviorally because of extraneous procedural or attentional factors, or because of a conflict between different brain processes (e.g., for maintaining the specific and rewarded learned pairing vs. generalizing to



| | | Congruity | Canonicity | Experiment | Congruity x canonicity | Congruity x experiment |
|---|---|---|---|---|---|---|
| ROIs Language | Temporal pole | 11.44 ** | <1 | 6.02 * | <1 | <1 |
| | anterior STS | 5.41 * | <1 | 42.31 *** | <1 | <1 |
| | posterior STS | 18.70 *** | 1.31 | 50.75 *** | <1 | 17.01 ** |
| | Temporo Parietal junction | 20.81 *** | 1.85 | 9.39 ** | <1 | <1 |
| | IFG orbitalis | 22.47 *** | <1 | 11.40 ** | <1 | 1.64 |
| | IFG triangularis | 16.98 *** | <1 | 22.42 *** | <1 | 10.45 * |
| | VWFA | 22.29 *** | <1 | 11.77 ** | <1 | <1 |
| ROIs Math | Left precentral BA44d | 29.71 *** | <1 | 4.1° | <1 | <1 |
| | Right precentral BA44d | 10.44 ** | 1.23 | <1 | 1.49 | <1 |
| | Left IPS | 27.4 *** | <1 | 1.81 | 1.77 | <1 |
| | Right IPS | 18.19 *** | 6.77 | 1.70 | 2.37 | 5.29 |
| | Left IT | 33.43 *** | <1 | 4.43° | <1 | <1 |
| | Right IT | 5.41 * | <1 | 7.76 * | <1 | <1 |
| | Left Cerebellum | 5.51 * | <1 | <1 | <1 | 2.87 |
| | Right Cerebellum | 19.20 *** | <1 | <1 | <1 | <1 |

**Table 5**

ROIs analyses: F-values of ANOVAs performed on the averaged betas of the main task across the 10% best voxels selected in an independent localizer in ROIs commonly activated in language and mathematical tasks. The language ROIs are presented as red areas on the sagittal (x=-50) and coronal (y=- 58) brain slices and the mathematical ROIs as yellow areas. The left white area corresponds to the VWFA; n=52; df=50; $p_{FDRcor}$: *** <0.001, ** < 0.01, *< 0.05, ° < 0.1.




## Figure 5.

### Summary of the two experiments in humans and monkeys.

(In experiment 1, $p_{voxel} < 0.001$ & $p_{cluster} < 0.05$ for humans and monkeys. In experiment 2, $p_{voxel}<0.005$ & cluster volume >50 in humans and $p_{voxel}<0.001$ & cluster volume >50 in monkeys.)

the reverse order). Here, we used fMRI and a passive paradigm to directly probe whether any area of the monkey brain would exhibit surprise at a violation of the reversal of a learned association. Our results show that this is not the case.

Interpretation must remain cautious, as there are also some occasional behavioral reports of spontaneous reversal of learned associations, for instance in one well-trained California sea lion and a Beluga whale (Kastak et al., 2001; Murayama et al., 2017; Schusterman and Kastak, 1998) and possibly in 1 out of 20 baboons in Medam et al. (2016). These studies may indicate that, with sufficient training, symbolic representation might eventually emerge in some animals, as also suggested by the small reversal trend in a recent behavior study in baboons (Chartier and Fagot, 2022). However, they may also merely show that animals may begin to spontaneously reverse new associations once they have received extensive training with bidirectional ones (Kojima, 1984). The bulk of the literature strongly suggests that while animals easily learn indexical associations, especially monkeys and chimpanzees (Diester and Nieder, 2007; Livingstone et al., 2010; Matsuzawa, 1985; Premack, 1971), but also dogs (Fugazza et al., 2021; Kaminski et al., 2004), vocal birds (e.g. Pepperberg, 2009) and even bees (Howard et al, 2019), they exhibit little or no evidence for genuine symbolic processing. Discriminating symbolic from indexical representations can be achieved by testing for spontaneous reversibility between the labels and the objects, as in the current study, or by testing for the presence of relationships among the labels (Nieder, 2009).

One previous study showed preliminary evidence for a lack of reversibility in macaque monkey inferotemporal cortex (Meyer & Olson, 2011), but only recorded on a subset of neurons, and after extensive training on pairs of visual images (816 exposures per pair). Interestingly, a similar set of arbitrary stimuli and extensive training protocol (258 trials per pair) was used in an fMRI study on stimulus association in humans, where congruity effects were also found to be restricted to early visual areas (Richter et al., 2018). It might have been that the extensive training lead to more low-level and rigid encoding in the trained direction. It is therefore instructive that, here, we found irreversibility after a very short training. Indeed, in experiment 2, just 24 exposures per pair were sufficient to observe a surprise effect in the canonical direction without generalization in the reverse direction -even after longer exposures. In addition, we strived to make the objects concrete and recognizable to the monkeys (by using pictures of toys that were familiar to them, taken from various angles), while the labels were as abstract as possible to promote a symbol-referent asymmetry in the pairs. We considered using macaque vocalizations for the sounds, but these already have a defined meaning, often emotional, that could have disrupted the experiments. Furthermore, the present animals had extensive experience with human speech. Finally, while the present lab setting could be judged artificial and not easily conducive to language acquisition, previous evidence indicates that human preverbal infants easily learn labels in such a setting (Mersad et al., 2021) and spontaneously reverse associations after only a short training period (Ekramnia and Dehaene-Lambertz, 2019; Kabdebon and Dehaene-Lambertz, 2019).

Non-human primates are often considered the animal model of choice to understand the neural correlates of high-level cognitive functions in humans (Feng et al., 2020; Newsome and Stein-Aviles, 1999; Roelfsema and Treue, 2014). Accordingly, many studies have emphasized the similarity between human and non-human primates in terms of brain anatomy, physiology and behavior (Caspari et al., 2018; De Valois et al., 1974; Erb et al., 2019; Hackett et al., 2001; Harwerth and Smith, 1985; Dante Mantini et al., 2012; D. Mantini et al., 2012; Mantini et al., 2011; Margulies et al., 2016; Petrides et al., 2012; Uhrig et al., 2014; Warren, 1974; Wilson et al., 2017; Wise, 2008). At the same time, important differences between human and monkey brains have been reported as well. Using a direct comparison with fMRI, some specific functional differences have been found (Denys et al., 2004a, 2004b; Mantini et al., 2013; Vanduffel et al., 2002). Particularly relevant is that in contrast to humans, monkeys show clear feature tuning in the prefrontal cortex, which is in line with the sensory activation we found in

monkey PFC (**Figure 1C**) and the involvement of monkey PFC in the congruity effect in experiment 2 (**Figure 4B**). Many anatomical differences have been reported between humans and monkeys using MRI as well as histological methods. In particular, the human brain is exceptionally large (Herculano-Houzel, 2012), and contains a number of structural differences compared to the brains of other primates (Chaplin et al., 2013; Leroy et al., 2015;

Neubert et al., 2014; Palomero-Gallagher and Zilles, 2019; Rilling, 2014; Schenker et al., 2010; Takemura et al., 2017). Notably, while the human arcuate fasciculus provides a strong direct connection between inferior prefrontal and temporal areas involved in language processing, this bundle is reduced and does not extend as anteriorly and as ventrally in other primates, including chimpanzees (Balezeau et al., 2020; Eichert et al., 2020; Rilling et al., 2012, 2008; Thiebaut de Schotten et al., 2012). Also, the PFC is selectively increased in terms of tissue volume (Chaplin et al., 2013; Donahue et al., 2018; Hill et al., 2010; Smaers et al., 2017). While this may not translate to a selective increase in terms of the number of PFC neurons (Gabi et al., 2016), dendritic arborizations and synaptic density are larger in human PFC (Elston, 2007; Hilgetag and Goulas, 2020; Shibata et al., 2021). These anatomical differences may underlie the fundamental differences in language learning abilities between these species, but this is still controversial (Hopkins et al., 2012; Iriki, 2006). Here, we show that reversibility of associations, a crucial element in the ability to attach symbols to objects and concepts, sharply differs between human and non-human primates and offers a more tractable way to investigate potential differences between species.

The areas that specifically activated in humans when the reversed association was violated were not limited to the classical language network in the left hemisphere. They extended bilaterally to homolog areas of the right hemisphere, which are involved for instance in the acquisition of musical languages (Patel, 2010). They also extend dorsally to the middle frontal gyrus and intraparietal sulcus which are involved in the acquisition of the language of numbers, geometry and higher mathematics (Amalric and Dehaene, 2016; Piazza, 2010; Wang et al., 2019). Finally, an ROI analysis shows that they also include the VWFA and vicinity. The VWFA is known to be sensitive to letters, but also to other visual symbols such as a new learned face-like script (Moore et al., 2014) or emblematic pictures of famous cities (e.g. the Eiffel tower for Paris; Song et al., 2012), and the nearby lateral inferotemporal cortex responds to Arabic numerals and other mathematical symbols (Amalric and Dehaene, 2016; Shum et al., 2013). Strikingly, these extended areas, shown in **Figure 2**, correspond to regions whose cortical expansion and connectivity patterns are maximally different in humans compared to other primates (Chaplin et al., 2013; Donahue et al., 2018; Hill et al., 2010; Smaers et al., 2017). They also fit with a previous fMRI comparison of humans and macaque monkeys, where humans were shown to exhibit uniquely abstract and integrative representations of numerical and sequence patterns in these regions (Wang et al., 2015).

In all of these studies, the observed changes are bilateral, extended, and go beyond the language network per se. Such an extended network does not fit with the hypothesis that a single localized system, such as natural language or a universal generative faculty, is the primary engine of all human-specific abstract symbolic abilities (Hauser and Watumull, 2017; Spelke, 2003). Rather, our results suggest that multiple parallel and partially dissociable human brain networks possess symbolic abilities and deploy them in different domains such as natural language, music and mathematics (Amalric and Dehaene, 2017; Chen et al., 2021; Dehaene et al., 2022; Fedorenko et al., 2011; Fedorenko and Varley, 2016).

The neurobiological mechanism that enables reversible symbol learning in humans remain to be discovered. Interestingly, most learning rules, such as spike-time-dependent plasticity, are sensitive to temporal order and timing, a feature of fundamental importance for predictive coding. In contrast, as indicated by the behavioral results of experiment 2, humans seem to forget the temporal order in which pairs of stimuli are presented when they store them at a symbolic

level. This has been interpreted as improper causal reasoning (Ogawa et al., 2010). Indeed, if A repeatedly precedes B, then perceiving A predict the appearance of B; but if B is observed, concluding to the likely presence of A is a logical fallacy. Still, brain mechanisms for temporal reversal do exist in the literature. The most prominent candidate, in both humans and non-human animals, is hippocampal-dependent neuronal replay of sequences of events, which can occur in both forward and reverse temporal order (Foster, 2017; Liu et al., 2019). Sequence reversal may be important during learning, in order to trace back to a memorized event that led to a reward. In line with this, a retroactive gradient has been shown in memory storage in humans, where memory is strongest for stimuli that were presented close to the reward but preceding it (Braun et al., 2018). This memory trace may explain the slight facilitation observed in baboons when they learn reversed congruent pairs relative to reversed incongruent pairs (Chartier et al, 2022). Although neuronal replay in both forward and reverse directions exists in non-human animals, it might be that this mechanism has selectively expanded to symbol-related areas of the human brain – a clear hypothesis for future work.

Obviously, even humans do not always disregard temporal order for all associations between stimulus pairs – for instance, they remember letters of the alphabet in a fixed temporal order (Klahr et al., 1983). Thus, future work should also clarify which conditions promote reversible symbolic learning. Here, the pairs comprised one fixed and abstract element (either linguistic or graphical), which served as a label, paired with several different views of a concrete object. In human infants, the association of a label with the presentation of objects helps them construct the object category, as revealed by several experiments in which infants discriminate between categories (Ferry et al., 2013), or correctly process the number of objects (Xu et al., 2005) when the categories and objects are named, but not in the absence of a label. Interestingly, preverbal infants are flexible and accept pictures as labels for a rule (Kabdebon et al, 2019), as well as monkey vocalizations and tones as labels for an animal category (Ferguson and Waxman, 2016; Ferry et al., 2013), whereas older infants who have been exposed to many social situations in which language is the primary symbolic medium to transfer information, expect symbolic labels to be in the native language (Perszyk and Waxman, 2019). Later, they recover flexibility suggesting that this transient limitation might be a contextual strategy due to the pivotal role of language in naming at this time of life.

While our results suggest a dramatic difference in the way human and non-human primates encode associations between sensory stimuli, several limitations of the present work should be kept in mind. First, due to ethical and financial reasons we only tested 4 monkeys, while we tested 55 humans in total. While it is common in primate physiological studies to report the results for 2 animals, this makes it challenging to extrapolate the results to a species of animals (Fries and Maris, 2022). To address this point, we combined the results from two different labs, collecting data from 2 animals in each lab. A second limitation is that we only tested macaque monkeys; non-human primates closer to humans, such as chimpanzees, might yield different conclusions, and chimpanzee Ai's failure of reversibility (Kojima, 1984), although striking, may not be representative. Similarly, reversible symbolic learning should be evaluated in vocal learners such as songbirds and parrots, as some demonstrate sophisticated flexible label learning (see e.g. Pepperberg and Carey, 2012). Furthermore, in dogs, social interactions between the dog and the experimenter during learning facilitate associations (Fugazza et al., 2021), as it is also the case in infants. Social cues were absent in our design, and whether they would favor a switch to a symbolic system might be interesting to explore. Finally, we only tested adult monkeys, yet there might be a critical period during which reversible symbolic representation might be possible with appropriate training procedures; indeed, juvenile macaques learn better and faster to associate an arbitrary label with visual quantities than adults (Srihasam et al., 2012). The present work provides a simple experimental paradigm that can easily be extended to all these cases, thus offering a unique opportunity to test whether humans are unique in their ability to acquire symbols.

# Methods

## Participants

We tested four adult rhesus macaques (male, 6-8 kg, 5-19 years of age). YS and JD participated in experiment 1 and JD, JC and DN in experiment 2. All procedures were conducted in accordance with the European convention for animal care (86-406) and the NIH's guide for the care and use of laboratory animals. They were approved by the Institutional Ethical Committee (CETEA protocol # 16-043) and by the ethical committee for animal research of the KU Leuven. Animal housing and handling were according to the recommendations of the Weatherall report, allowing extensive locomotor behavior, social interactions, and foraging. All animals were group-housed (cage size at least 16-32 m3) with diverse cage enrichment (auditory and visual stimuli, toys, foraging devices etc.).

We also tested 55 healthy human subjects with no known neurological or psychiatric pathology (Exp. 1, n=31; Exp2., n=24; in experiment 2, an additional 3 subjects were not included because they showed no evidence of learning the canonical pairs). Human subjects gave written informed consent to participate in this study, which was approved by the French national Ethics Committee.

## Stimuli

Five sets of images were used (**Supplementary Figure 1**). The two first sets were 3D renderings of objects differing in their visual properties and semantic categories. As they might be considered as more familiar to humans, the other three sets of objects were photographs of monkey toys which the monkeys were exposed to in their home cages for at least 2 weeks prior to the training blocks. They were mostly geometrical 3D objects with no evident and consistent name for naive human participants. The rendering and photos were taken from 8 different viewpoints. These stimuli were used in both experiments and are called "object" thereafter.

A label was associated to each object in each set. For experiment 1, the labels were auditory French pseudo- words with large differences in the number and identity of their syllables within each set (e.g. "tøj$^{\alpha^{\circ}}$", "gliʃu","byɲyɲy", "kʁɛfila"). Note that monkeys were daily exposed to French radio and television as well as to French-speaking animal caretakers. or experiment 2, the labels were abstract black-and-white shapes, difficult to name and similar to the lexigrams used to train chimpanzees to communicate with humans (Matsuzawa, 1985) .

## Experimental paradigm

### Stimulus presentation

Each set to be learned comprised 4 pairs. Two pairs were presented in the label- object direction (L1-O1 & L3-O3), and two in the object-label direction (O2-L2 & O4-L4). Labels were speech sounds in experiment 1, and black-and-white shapes in experiment 2. In each trial, the first stimulus (label or object) was presented during 700ms, followed by an inter-stimulus-interval of 100ms then the second stimulus during 700ms. The pairs were separated by a variable inter-trial-interval of 3-5 seconds. The visual stimuli were ~8 degrees in diameter, centered on the screen, with an average luminance set equal to the background. At each trial, the orientation of the object was randomly chosen among the 8 possibilities. A cross was present at the center of the screen when no visual stimulus was present. Auditory stimuli were presented to both ears at 80dB.

## Training

The experiment was designed to be also tested in 3-month-old human infants (Ekramnia et al, in preparation), which explains our choice of short training sessions over 3 consecutive days because of the short attention span in infants and the reported benefit of sleep for encoding word meaning after a learning session (Friedrich et al., 2017). Therefore, training consisted of three short videos presenting 24 trials as described above (one video for each of the 3 training days). Two pairs (one in each direction) were introduced on the first day of training (e.g., L1-O1 and O2-L2). First, one pair was shown for 6 trials, then the other pair for 6 trials, then the two pairs were randomly presented for 6 trials each. On the second day of training, the two other pairs (L3-O3 and O4-L4) were presented using the same procedure than on day 1. On the third day, all pairs were randomly presented (6 presentations each). The object-label pairing was constant but the direction of presentation (O-L or L-O) and the introduction of the pair on the first or second day was counterbalanced across participants.

## Human protocol

n experiment 1, the participants came to the lab to watch the first video, and on the next two consecutive days they received a web link on which the two videos were uploaded for each day. For experiment 2, all 3 videos were sent via a web link. The participants were only instructed to look attentively at each movie (24 trials, ~3 min long) one time on a given day. The participants came for the fMRI session on the fourth day. Each participant saw only one set of objects-labels, either stimulus set 2 or stimulus set 3, distributed equally across participants.

In experiment 2, we added a behavioral test at the end of the MRI session to check their learning. They were shown all 16 possible trial pairs (incongruent and congruent in canonical and non-canonical order), plus 16 never seen, one by one. For each of them, they were asked to rate how frequently they had seen them (on a 5-level scale ranging from never to rarely, sometimes, often and always). The results were analyzed using a 5-level ANOVA which included the canonicity X congruity 2x2 design. A computer crash erased responses from two participants and one subject did not participate leaving 21 subjects for this analysis.

## Monkey protocol

Monkeys were implanted with an MR-compatible headpost under general anesthesia. The animals were first habituated to remain calm in a chair inside a mock MRI setup, and trained to fixate a small dot (0.25 degrees) within a virtual window of 1.25-2 degrees diameter (Uhrig et al., 2014). Then similar to the human participants, they received 1 training block per day for 3 consecutive days (24 trials per block) for each stimulus set. Rewards were given at regular intervals, asynchronous with the visual and auditory stimulus presentation. On the fourth day, they were scanned while being presented with the test blocks for the corresponding stimulus set. All monkeys were trained and tested with all of the stimulus sets.

For experiment 1, after the first imaging session at day 4 which did not show learning (no difference between congruent and incongruent pairs in the canonical direction), monkeys were further trained for an additional 2 weeks (~80 blocks), and then scanned each day during 4 days. Then a new set of four object- label pairs was presented with the same training and testing design. So, training and testing took 3 consecutive weeks for each of the five stimulus sets.

In experiment 2, a reward was introduced to promote monkeys' engagement in the task. The amount of reward that the monkeys received after successfully fixating throughout the pair presentation was either increased or decreased for a duration of 1450ms (starting 100ms after the offset of the second stimulus), depending on the identity of the visual object. The amount of reward remained the same, but the time in between consecutive rewards was set either twice as

short (for high rewards) or twice as long (for low rewards). For each direction, one visual object was associated with a high reward while the other one was associated with a low reward (see **Supplementary Figure 2**). By design, the two pairs that were averaged for each of the critical tested dimension (direction, congruity and canonicity of the pair) had opposite reward size, making reward size an orthogonal design element. The first stimulus set was used for procedural training on this reward association paradigm for 2 weeks. Stimulus sets 2-5 were used for training as in experiment 1 (with 1 block per day for 3 consecutive days) and an fMRI test session on the fourth day.

### Test in MRI

The MRI session comprised 4 test blocks in humans and between 12 and 32 blocks in monkeys. In both humans and monkeys, each block started with 4 trials in the learned direction (congruent canonical trials), one trial for each of the 4 pairs (2 O-L and 2 L-O pairs). The rest of the block consisted of 40 trials in which 70% of trials were identical to the training; 10% were incongruent pairs but the direction (O-L or L-O) was correct (incongruent canonical trials), thus testing whether the association was learned; 10% were congruent pairs but the direction within the pairs was reversed relative to the learned pairs (congruent reversed trials) and 10% were incongruent pairs in reverse (incongruent reversed trials). As the percentage of congruent and incongruent pairs was the same in the reversed direction, a difference can only be due to a generalization from the canonical direction. For incongruent trials, the incongruent stimulus always came from the pair presented in the same direction (see **figure 1**), in order to avoid that a change of position within the pair itself (1st or 2nd stimulus) induced the perception of an incongruity.

Human participants were only instructed to keep their eyes fixed on the fixation point and pay attention to the stimuli. The monkeys were rewarded for keeping their eyes fixed on the fixation point, as in the training. In Experiment 1, the reward was constant, whereas in Experiment 2, they received the differential reward that was implemented during training.

## Data acquisition

For experiment 1, both humans and monkeys were scanned with the 3T Siemens Prisma at NeuroSpin using a T2*-weighted gradient echo-planar imaging (EPI) sequence, using a 64-channel head coil for humans and a customized eight-channel phased-array surface coil (KU Leuven, Belgium) for monkeys. The imaging parameters were the following: in humans, resolution: 1.75mm isotropic, TR: 1.81s, TE: 30.4ms, PF: 7/8, MB3, slices: 69; in monkeys, resolution: 1.5mm isotropic, TR: 1.08s, TE: 13.8ms, PF: 6/8, iPAT2, slices: 34.

Monkeys were trained to sit in a sphinx position in a primate chair with their head fixed. MION (monocrystalline iron oxide nanoparticle, Molday Ion, BioPAL, Worchester MA) contrast agent (10 mg/kg, i.v.) was injected to monkeys before scanning (Vanduffel et al., 2001). Eye movements were monitored and recorded by an eye tracking system (EyeLink 1000, SR Research, Ottawa, Canada). In total, we recorded 583 valid runs, 278 for YS and 305 for JD.

For Experiment 2, the settings remained the same for the humans and for one of the monkeys (JD). Two new monkeys (JC and DN) were included at the Laboratory of Neuro- and Psychophysiology of KU Leuven and scanned with a 3T Siemens Prisma using a T2*-weighted gradient echo-planar imaging (EPI) sequence. For JC, an external 8-channel coil was used and the imaging parameters were the following: resolution: 1.25mm isotropic, TR: 0.9s, T7: 15ms, PF: 6/8, iPAT3, multi-band 2, slices: 52. For DN, an implanted 8- channel coil was used and the imaging parameters were the following: resolution: 1.25mm isotropic, TR: 0.9s, TE: 15ms, PF: 6/8, iPAT3, multi-band 2, slices: 40. Monkeys were trained to sit in a sphinx position in a primate chair with their head fixed, and MION was again injected to before scanning (11 mg/kg, i.v.). Eye movements were monitored and recorded by an eye tracking system (ETL200, ISCAN inc., Woburn, MA, USA). The animals were

also required to keep their hands in a box in front of the chair (as verified with optical sensors), which limited body motion. In total, we recorded 279 valid runs, 81 for JD, 106 for JC and 92 for DN.

## Preprocessing of monkey fMRI data

Functional images were reoriented, realigned, resampled (1.00 mm isotropic) and coregistered to the anatomical template of the monkey Montreal Neurologic Institute (Montreal, Canada) space using Pypreclin, a custom-made scripts of Python programming language (Tasserie et al., 2020).

Eye-data was inspected for each run for quality. Only runs with more than 85% fixation (virtual window of 2-2.5 degrees diameter) were included for further analyses (n=16 excluded in experiment 1 and n=14 excluded in experiment 2). Moreover, a trial was excluded if the eyes were closed for more than 650ms (out of 700) while an image was present on the screen. In experiment 1, the top 5% of runs were motion was strongest across monkeys were excluded (n=30) because there remained significant residual motion. In total, for experiment 1, 395 runs remained to be analyzed, 184 for YS and 211 for JD. For experiment 2, 268 runs remained, 77 for JD, 107 for JC and 84 for DN.

## Preprocessing of human fMRI data

SPM12 (*http://www.fil.ion.ucl.ac.uk/spm*) was used for preprocessing of human data as well as first and second level models. Preprocessing consisted of standard preprocessing pipeline, including slice-time correction, realign, top-up correction, segmentation, normalization to standard MNI space and smoothing with a 4-mm isotropic Gaussian.

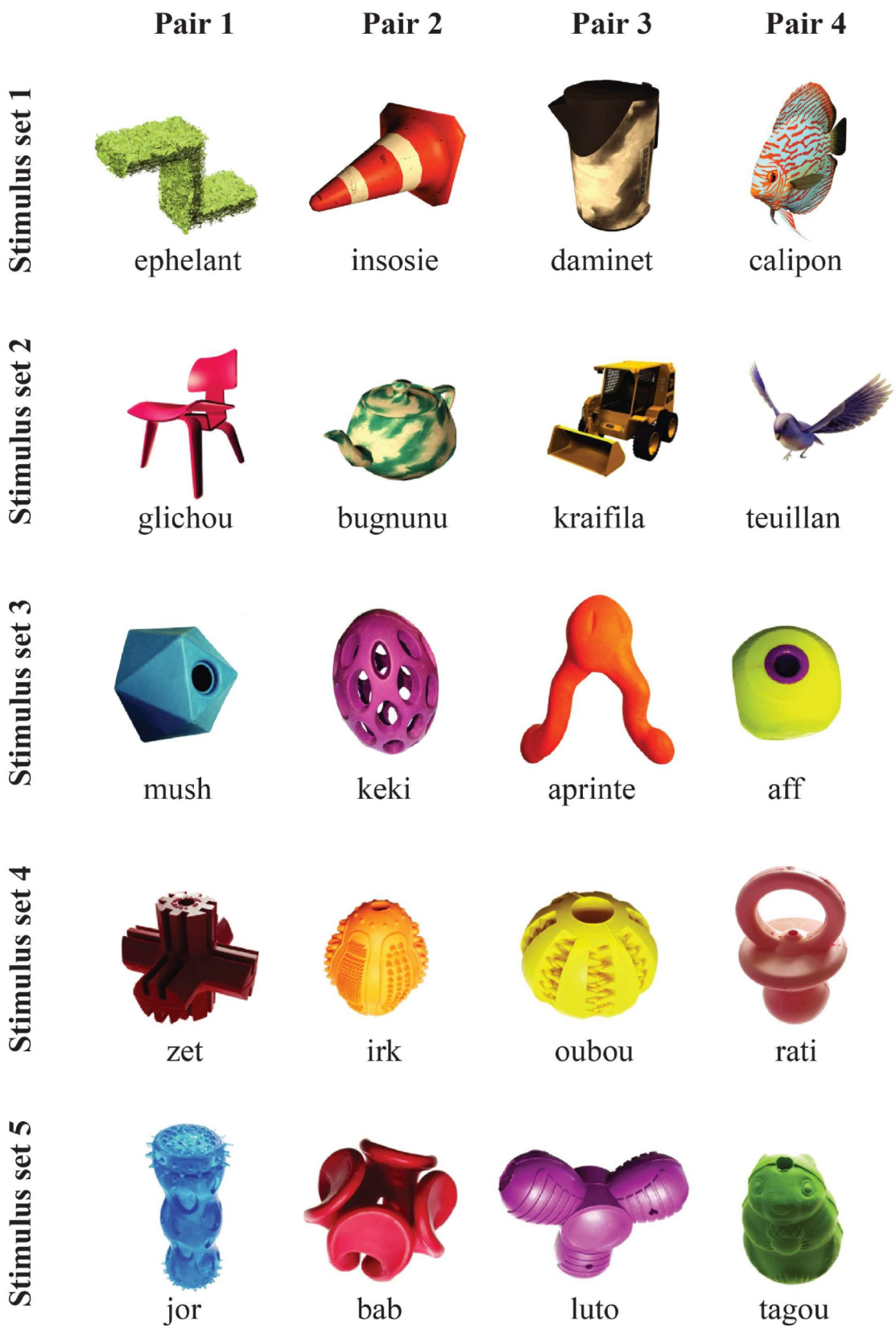
## First and second-level analyses

After imaging preprocessing, active brain regions were identified by performing voxel-wise GLM analyses implemented in SPM12 in both monkeys and humans. For the first experiment, in a first-level SPM model, the twelve predictors included: (1-4) the onsets of the first stimulus of the pair (4 regressors consisting in the combinations of audio/visual and canonical/non-canonical factors), and (5-12) the onsets of the second stimulus (8 regressors consisting in the combinations of audio/visual, canonical/non-canonical and congruent/incongruent factors). These twelve events were modeled as delta functions convolved with the canonical hemodynamic response function (for MION in case of monkeys). Parameters of head motion derived from realignment were also included in the model as covariates of no interest. Contrast images for the effect of congruity (incongruent-congruent canonical and incongruent - congruent non-canonical) as well as the interaction (congruity x canonicity) were computed. For the second experiment, the analysis was the same, except that given the two elements of the pair were in the same visual modality, only a single predictor was used for each stimulus pair, giving 4 predictors: the onsets of the second stimulus of the pair, with congruent/incongruent and canonical/non-canonical as the two factors. For the monkeys, an additional factor was whether the pair was associated with a high or a low reward, giving 8 predictors. In this case, the temporal derivative of the hemodynamic response function was added to the model as well. Before entering the second-level analysis, the data was smoothed again, using a 5mm smoothing kernel in humans and 2 mm in monkeys.

For the second-level group analysis, subjects were taken as the statistical unit for the humans and runs were taken as statistical units for the monkeys. One-sample t-tests were performed on the contrast images to test for the effect of the condition. Results are reported at an uncorrected voxelwise threshold of $p<0.001$ and a cluster $p<0.05$ corrected for multiple comparisons (FDR).

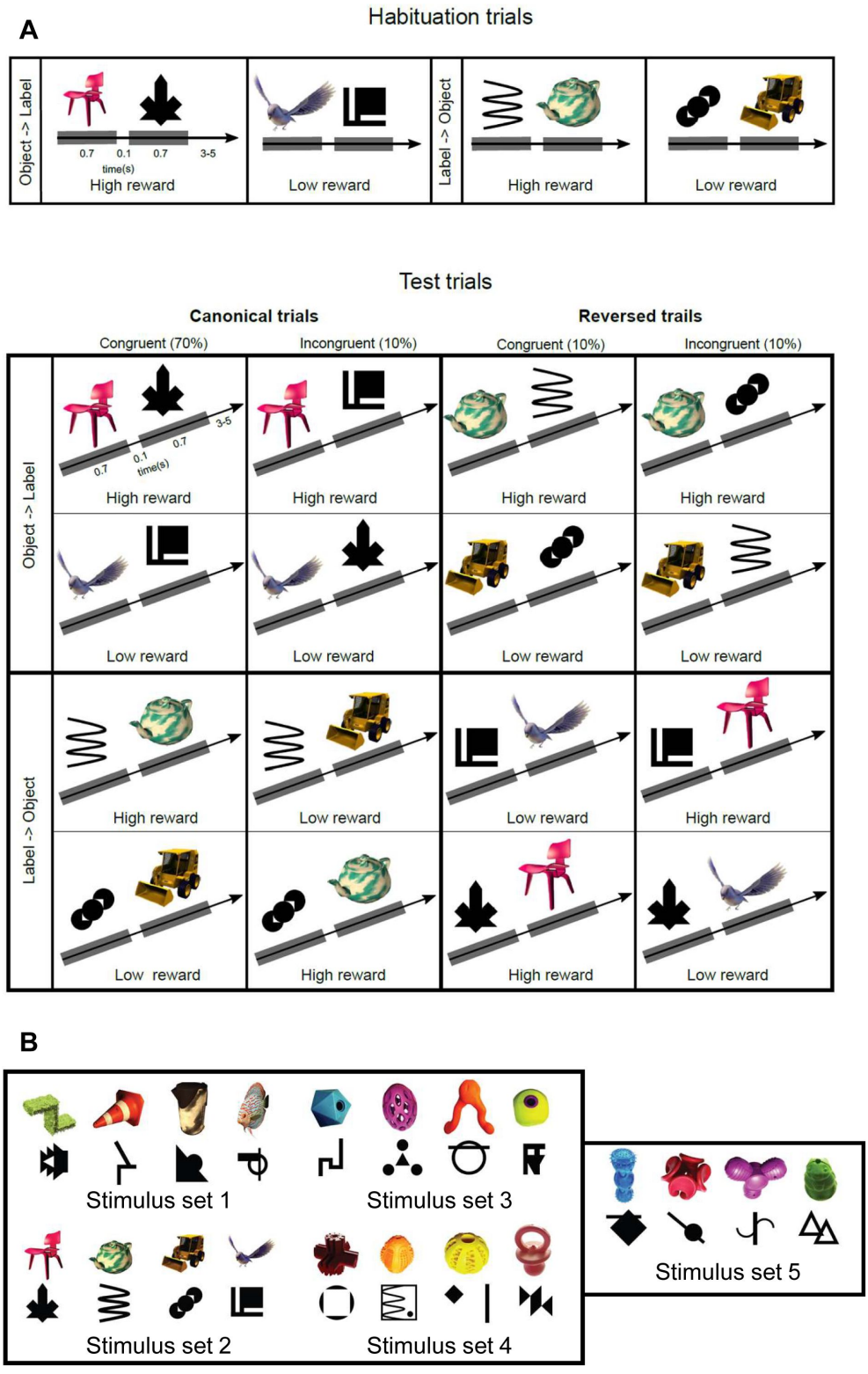
## ROI analyses

In a separate localizer, human participants listened and read short sentences. In some of the sentences, the participants were asked to compute easy mathematical operations (math sentences). Subtracting activations to math and non-math sentences allowed to separate the regions more involved in mathematical cognition than in general sentence comprehension and reciprocally. We selected seven left-hemispheric regions previously reported as showing a language-related activation (Pallier et al., 2011), 6 bilateral ROIs showing mathematically-related activations (Amalric & Dehaene, 2016) and finally a 10-radius sphere around the VWFA [-45 -57 -12]. In these ROIs, we recovered the coordinates of each participant's 10% best voxels in the comparisons: sentences vs rest for the 6 language Rois plus reading vs listening for the VWFA, and numerical vs non-numerical sentences for the 8 mathematical ROIs. We extracted the beta of these voxels and performed ANOVAs with Congruity and Canonicity as within-subject factors and experiment as between-subjects factor. " Two participants in experiment 1 and one in experiment 2 had no localizer, leaving 52 participants (n=29 and n= 23) for these analyses. P-values were FDR corrected considering all 15 ROIs in each comparison.



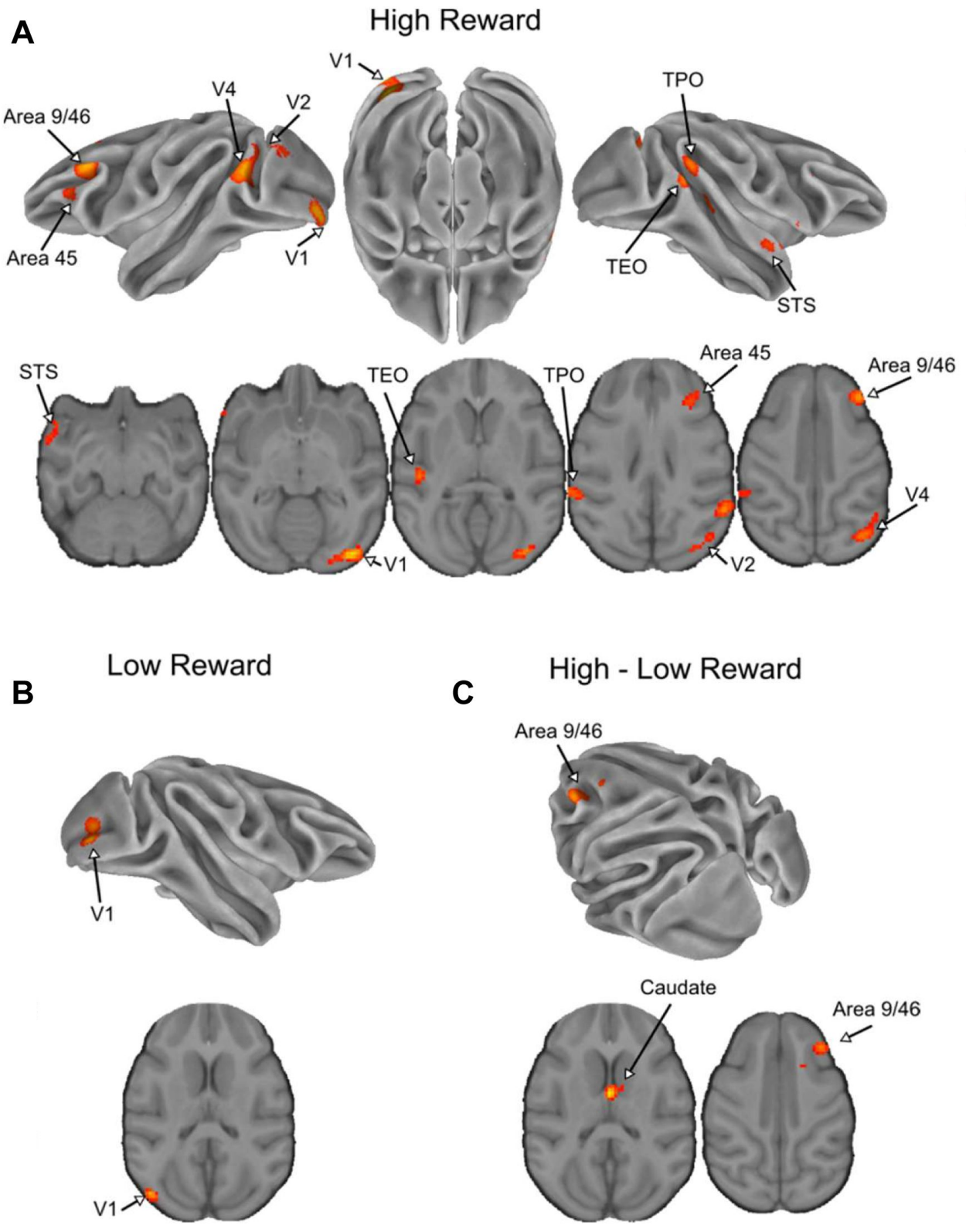

**Supplementary figure 1**

**Stimulus sets for experiment 1.**




## Supplementary Figure 2.

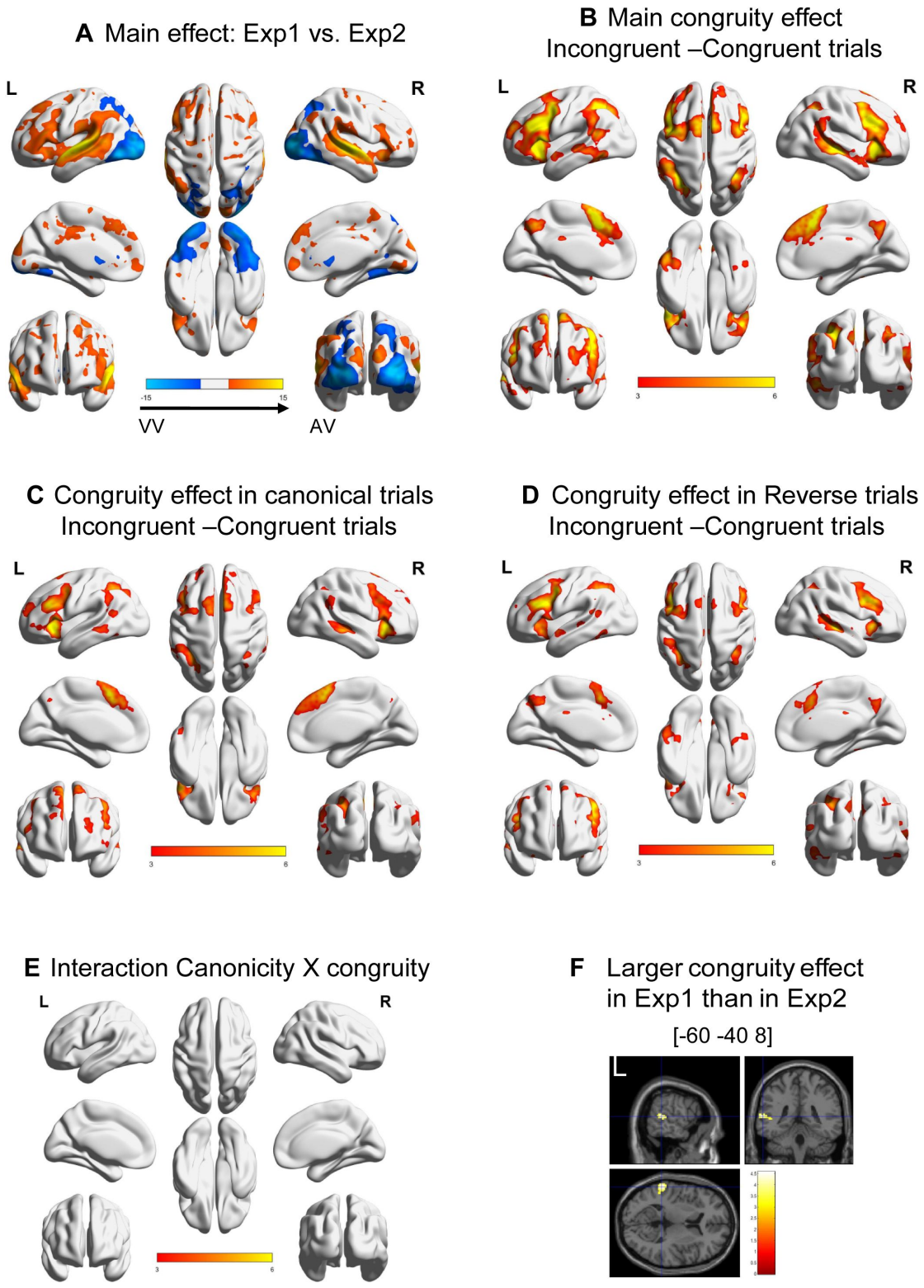
**A) Complete description of the task paradigm for visual-visual label learning.** Subjects were habituated to 4 different visual-visual pairs during three days. Two pairs were in the 'object-label' order and two pairs in the 'label-object' order. During the test phase, the same canonical order was kept in 80% of the trials, including 10% of incongruent pairs. In reversed trials (20% of trials), the pairs were either congruent (10%) or incongruent (10%) with the learning. For the monkeys, one pair in each direction was associated with a high reward while the other one was associated with a low reward, making the reward size orthogonal to congruity and canonicity. **B)** Stimulus sets for experiment 2 in monkeys. Humans were tested with stimulus set 2.




## Supplementary Figure 3

**Effect of reward for the visual-visual task in non-human primates.**

**A**) Significant clusters from the incongruent-congruent canonical contrast in low reward trials. **B**) Significant from the incongruent-congruent canonical contrast in high reward trials. **C**) Significant clusters from the interaction between congruity and reward. $p_{voxel}$<0.001 & $p_{cluster}$ <0.05 in all panels.




**Supplementary Figure 4.**

**Analyses of all human participants in experiments 1 and 2 merged.**

**A)** Main effect of experiment. **B)** Main effect of congruity, **C)** Effect of congruity in the canonical trials and D) in the reversed trials. **E)** No significant cluster was observed for the interaction canonicity X congruity. **F)** slices in the 3 planes showing the only significant cluster in the Experiment X Congruity interaction. $p_{voxel}$<0.001 & $p_{cluster}$ <0.05 in all panels

# References


Amalric M, Dehaene S (2017) **Cortical circuits for mathematical knowledge: evidence for a major subdivision within the brain's semantic networks** *Philos Trans R Soc Lond, B, Biol Sci* **373** https://doi.org/10.1098/rstb.2016.0515

Amalric M, Dehaene S (2016) **Origins of the brain networks for advanced mathematics in expert mathematicians** *ProcNatlAcadSciUSA* **113**:4909–4917

Balezeau F, Wilson B, Gallardo G, Dick F, Hopkins W, Anwander A, Friederici AD, Griffiths TD, Petkov CI (2020) **Primate auditory prototype in the evolution of the arcuate fasciculus** *Nat Neurosci* **23**:611–614 https://doi.org/10.1038/s41593-020-0623-9

Berwick RC, Chomsky N (2016) **Why Only Us: Language and Evolution**

Braun EK, Wimmer GE, Shohamy D (2018) **Retroactive and graded prioritization of memory by reward** *Nat Commun* **9** https://doi.org/10.1038/s41467-018-07280-0

Caspari N, Arsenault JT, Vandenberghe R, Vanduffel W (2018) **Functional Similarity of Medial Superior Parietal Areas for Shift-Selective Attention Signals in Humans and Monkeys** *Cerebral Cortex* **28**:2085–2099 https://doi.org/10.1093/cercor/bhx114

Chaplin TA, Yu H-H, Soares JGM, Gattass R, Rosa MGP (2013) **A Conserved Pattern of Differential Expansion of Cortical Areas in Simian Primates** *Journal of Neuroscience* **33**:15120–15125 https://doi.org/10.1523/JNEUROSCI.2909-13.2013

Chartier TF, Fagot J (2022) **Simultaneous learning of directional and non-directional stimulus relations in baboons (Papio papio)** *Learn Behav* https://doi.org/10.3758/s13420-022-00522-8

Chen X, Affourtit J, Ryskin R, Regev TI, Norman-Haignere S, Jouravlev O, Malik-Moraleda S, Kean H, Varley R, Fedorenko E (2021) **The human language system does not support music processing** https://doi.org/10.1101/2021.06.01.446439

De Valois RL, Morgan H, Snodderly DM. (1974) **Psychophysical studies of monkey Vision-III. Spatial luminance contrast sensitivity tests of macaque and human observers** *Vision Research* **14**:75–81 https://doi.org/10.1016/0042-6989(74)90118-7

Deacon TW (1998) **The Symbolic Species – The Co–evolution of Language & the Brain**

Dehaene S, Al Roumi F, Lakretz Y, Planton S, Sablé-Meyer M. (2022) **Symbols and mental programs: a hypothesis about human singularity** *Trends in Cognitive Sciences* **26**:751–766 https://doi.org/10.1016/j.tics.2022.06.010

Dehaene S, Meyniel F, Wacongne C, Wang L, Pallier C (2015) **The Neural Representation of Sequences: From Transition Probabilities to Algebraic Patterns and Linguistic Trees** *Neuron* **88**:2–19 https://doi.org/10.1016/j.neuron.2015.09.019

Denys K, Vanduffel W, Fize D, Nelissen K, Peuskens H, van ED, Orban GA (2004) **The processing of visual shape in the cerebral cortex of human and nonhuman primates: a functional magnetic resonance imaging study** *JNeurosci* **24**:2551–2565

Denys K, Vanduffel W, Fize D, Nelissen K, Sawamura H, Georgieva S, Vogels R, van ED, Orban GA (2004) **Visual activation in prefrontal cortex is stronger in monkeys than in humans** *JCogn Neurosci* **16**:1505–1516

Diester I, Nieder A (2007) **Semantic associations between signs and numerical categories in the prefrontal cortex** *PLoS Biol* **5** https://doi.org/10.1371/journal.pbio.0050294

Donahue CJ, Glasser MF, Preuss TM, Rilling JK, Essen DCV (2018) **Quantitative assessment of prefrontal cortex in humans relative to nonhuman primates** *PNAS* **115**:E5183–E5192 https://doi.org/10.1073/pnas.1721653115

Duncan J (2010) **The multiple-demand (MD) system of the primate brain: mental programs for intelligent behaviour** *Trends in Cognitive Sciences* **14**:172–179 https://doi.org/10.1016/j.tics.2010.01.004

Eichert N, Robinson EC, Bryant KL, Jbabdi S, Jenkinson M, Li L, Krug K, Watkins KE, Mars RB (2020) **Cross-species cortical alignment identifies different types of anatomical reorganization in the primate temporal lobe** *eLife* **9** https://doi.org/10.7554/eLife.53232

Ekramnia M, Dehaene-Lambertz G. (2019) **Investigating bidirectionality of associations in young infants as an approach to the symbolic system** *Presented at the CogSci.*

Elston GN, Kaas JH (2007) **4.13 - Specialization of the Neocortical Pyramidal Cell during Primate Evolution** *Evolution of Nervous Systems* :191–242 https://doi.org/10.1016/B0-12-370878-8/00164-6

Erb J, Armendariz M, De Martino F, Goebel R, Vanduffel W, Formisano E. (2019) **Homology and Specificity of Natural Sound-Encoding in Human and Monkey Auditory Cortex** *Cerebral Cortex* **29**:3636–3650 https://doi.org/10.1093/cercor/bhy243

Fedorenko E, Behr MK, Kanwisher N (2011) **Functional specificity for high-level linguistic processing in the human brain** *Proc Natl Acad Sci U S A* **108**:16428–33 https://doi.org/10.1073/pnas.1112937108

Fedorenko E, Duncan J, Kanwisher N (2013) **Broad domain generality in focal regions of frontal and parietal cortex** *Proc Natl Acad Sci USA* **110**:16616–16621 https://doi.org/10.1073/pnas.1315235110

Fedorenko E, Varley R (2016) **Language and thought are not the same thing: evidence from neuroimaging and neurological patients** *Ann N Y Acad Sci* **1369**:132–153 https://doi.org/10.1111/nyas.13046

Felleman DJ, Van Essen DC. (1991) **Distributed hierarchical processing in the primate cerebral cortex** *CerebCortex* **1**:1–47

Feng G, Jensen FE, Greely HT, Okano H, Treue S, Roberts AC, Fox JG, Caddick S, Poo M, Newsome WT, Morrison JH (2020) **Opportunities and limitations of genetically modified nonhuman primate models for neuroscience research** *PNAS* **117**:24022–24031 https://doi.org/10.1073/pnas.2006515117

Ferguson B, Waxman SR (2016) **What the [beep]? Six-month-olds link novel communicative signals to meaning** *Cognition* **146**:185–9 https://doi.org/10.1016/j.cognition.2015.09.020

Ferry AL, Hespos SJ, Waxman SR (2013) **Nonhuman primate vocalizations support categorization in very young human infants** *Proc Natl Acad Sci U S A* **110**:15231–5 https://doi.org/10.1073/pnas.1221166110

Fitch WT, Hauser MD, Chomsky N (2005) **The evolution of the language faculty: clarifications and implications** *Cognition* **97**:179–210

Foster DJ (2017) **Replay Comes of Age** *Annual Review of Neuroscience* **40**:581–602 https://doi.org/10.1146/annurev-neuro-072116-031538

Friedrich M, Wilhelm I, Molle M, Born J, Friederici AD (2017) **The Sleeping Infant Brain Anticipates Development** *Curr Biol* **27**:2374–2380 https://doi.org/10.1016/j.cub.2017.06.070

Fries P, Maris E (2022) **What to Do If N Is Two?** *Journal of Cognitive Neuroscience* **34**:1114–1118 https://doi.org/10.1162/jocn_a_01857

Fugazza C, Andics A, Magyari L, Dror S, Zempléni A, Miklósi Á (2021) **Rapid learning of object names in dogs** *Sci Rep* **11** https://doi.org/10.1038/s41598-021-81699-2

Gabi M, Neves K, Masseron C, Ribeiro PFM, Ventura-Antunes L, Torres L, Mota B, Kaas JH, Herculano- Houzel S (2016) **No relative expansion of the number of prefrontal neurons in primate and human evolution** *Proceedings of the National Academy of Sciences* **113**:9617–9622 https://doi.org/10.1073/pnas.1610178113

Ghazizadeh A, Griggs W, Leopold DA, Hikosaka O (2018) **Temporal–prefrontal cortical network for discrimination of valuable objects in long-term memory** *PNAS* **115**:E2135–E2144 https://doi.org/10.1073/pnas.1707695115

Hackett TA, Preuss TM, Kaas JH (2001) **Architectonic identification of the core region in auditory cortex of macaques, chimpanzees, and humans** *J Comp Neurol* **441**:197–222 https://doi.org/10.1002/cne.1407

Harwerth RS, Smith EL (1985) **Rhesus monkey as a model for normal vision of humans** *Am J Optom Physiol Opt* **62**:633–641 https://doi.org/10.1097/00006324-198509000-00009

Hauser MD, Chomsky N, Fitch WT (2002) **The faculty of language: what is it, who has it, and how did it evolve?** *Science* **298**:1569–1579

Hauser MD, Watumull J (2017) **The Universal Generative Faculty: The source of our expressive power in language, mathematics, morality, and music** *Journal of Neurolinguistics* https://doi.org/10.1016/j.jneuroling.2016.10.005

Henshilwood CS, d'Errico F, Yates R, Jacobs Z, Tribolo C, Duller GAT, Mercier N, Sealy JC, Valladas H, Watts I, Wintle AG. (2002) **Emergence of Modern Human Behavior: Middle Stone Age Engravings from South Africa** *Science* **295**:1278–1280 https://doi.org/10.1126/science.1067575

Herculano-Houzel S (2012) **The remarkable, yet not extraordinary, human brain as a scaled-up primate brain and its associated cost** *PNAS* **109**:10661–10668 https://doi.org/10.1073/pnas.1201895109

Hilgetag CC, Goulas A (2020) **'Hierarchy' in the organization of brain networks** *Philosophical Transactions of the Royal Society B: Biological Sciences* **375** https://doi.org/10.1098/rstb.2019.0319

Hill J, Inder T, Neil J, Dierker D, Harwell J, Van Essen D. (2010) **Similar patterns of cortical expansion during human development and evolution** *Proceedings of the National Academy of Sciences* **107**:13135–13140 https://doi.org/10.1073/pnas.1001229107

Hopkins WD, Russell JL, Schaeffer JA (2012) **The neural and cognitive correlates of aimed throwing in chimpanzees: a magnetic resonance image and behavioural study on a unique form of social tool use** *Philos Trans R Soc Lond B Biol Sci* **367**:37–47 https://doi.org/10.1098/rstb.2011.0195

Howard SR, Avarguès-Weber A, Garcia JE, Greentree AD, Dyer AG (2019) **Symbolic representation of numerosity by honeybees (Apis mellifera): matching characters to small quantities** *Proceedings of the Royal Society B: Biological Sciences* **286** https://doi.org/10.1098/rspb.2019.0238

Imai M, Murai C, Miyazaki M, Okada H, Tomonaga M (2021) **The contingency symmetry bias (affirming the consequent fallacy) as a prerequisite for word learning: A comparative study of pre-linguistic human infants and chimpanzees** *Cognition* **214** https://doi.org/10.1016/j.cognition.2021.104755

Iriki A (2006) **The neural origins and implications of imitation, mirror neurons and tool use. Current Opinion in Neurobiology** *Motor systems / Neurobiology of behaviour* **16**:660–667 https://doi.org/10.1016/j.conb.2006.10.008

Kabdebon C, Dehaene-Lambertz G (2019) **Symbolic labeling in 5-month-old human infants** *PNAS* **116**:5805–5810 https://doi.org/10.1073/pnas.1809144116

Kaminski J, Call J, Fischer J (2004) **Word learning in a domestic dog: evidence for "fast mapping."** *Science* **304**:1682–3 https://doi.org/10.1126/science.1097859

Kaposvari P, Kumar S, Vogels R (2018) **Statistical Learning Signals in Macaque Inferior Temporal Cortex** *Cereb Cortex* **28**:250–266 https://doi.org/10.1093/cercor/bhw374

Kastak CR, Schusterman RJ, Kastak D (2001) **Equivalence classification by California sea lions using class- specific reinforcers** *J Exp Anal Behav* **76**:131–158 https://doi.org/10.1901/jeab.2001.76-131

Kietzmann C, Keil G, Kreft N (2019) **Aristotle on the Definition of What It Is to Be Human** *Aristotle's Anthropology* :25–43 https://doi.org/10.1017/9781108131643.002

Klahr D, Chase WG, Lovelace EA (1983) **Structure and process in alphabetic retrieval** *Journal of Experimental Psychology: Learning, Memory, and Cognition* **9**:462–477 https://doi.org/10.1037/0278-7393.9.3.462

Kojima T (1984) **Generalization between productive use and receptive discrimination of names in an artificial visual language by a chimpanzee** *Int J Primatol* **5**:161–182 https://doi.org/10.1007/BF02735739

Leroy F, Cai Q, Bogart SL, Dubois J, Coulon O, Monzalvo K, Fischer C, Glasel H, Van der Haegen L, Bénézit A, Lin C-P, Kennedy DN, Ihara AS, Hertz-Pannier L, Moutard M-L, Poupon C, Brysbaert M, Roberts N, Hopkins WD, Mangin J-F, Dehaene-Lambertz G. (2015) **New human-specific brain landmark: The depth asymmetry of superior temporal sulcus** *Proceedings of the National Academy of Sciences* **112**:1208–1213 https://doi.org/10.1073/pnas.1412389112

Lipkens R, Kop PFM, Matthijs W (1988) **A test of symmetry and transitivity in the conditional discrimination performances of pigeons** *J Exp Anal Behav* **49**:395–409 https://doi.org/10.1901/jeab.1988.49-395

Liu Y, Dolan RJ, Kurth-Nelson Z, Behrens TEJ (2019) **Human Replay Spontaneously Reorganizes Experience** *Cell* **178**:640–652 https://doi.org/10.1016/j.cell.2019.06.012

Livingstone MS, Pettine WW, Srihasam K, Moore B, Morocz IA, Lee D (2014) **Symbol addition by monkeys provides evidence for normalized quantity coding** *Proc Natl Acad Sci U S A* **111**:6822–7 https://doi.org/10.1073/pnas.1404208111

Livingstone MS, Srihasam K, Morocz IA (2010) **The benefit of symbols: monkeys show linear, human- like, accuracy when using symbols to represent scalar value** *Anim Cogn* **13**:711–9 https://doi.org/10.1007/s10071-010-0321-1

Mantini D, Corbetta M, Romani GL, Orban GA, Vanduffel W (2013) **Evolutionarily novel functional networks in the human brain?** *JNeurosci* **33**:3259–3275

Mantini D, Corbetta M, Romani GL, Orban GA, Vanduffel W (2012) **Data-driven analysis of analogous brain networks in monkeys and humans during natural vision** *NeuroImage* **63**:1107–1118 https://doi.org/10.1016/j.neuroimage.2012.08.042

Mantini D, Gerits A, Nelissen K, Durand J-B, Joly O, Simone L, Sawamura H, Wardak C, Orban GA, Buckner RL, Vanduffel W (2011) **Default mode of brain function in monkeys** *J Neurosci* **31**:12954–12962 https://doi.org/10.1523/JNEUROSCI.2318-11.2011

Mantini D., Hasson U, Betti V, Perrucci MG, Romani GL, Corbetta M, Orban GA, Vanduffel W (2012) **Interspecies activity correlations reveal functional correspondence between monkey and human brain areas** *NatMethods* **9**:277–282

Margulies DS, Ghosh SS, Goulas A, Falkiewicz M, Huntenburg JM, Langs G, Bezgin G, Eickhoff SB, Castellanos FX, Petrides M, Jefferies E, Smallwood J (2016) **Situating the default-mode network along a principal gradient of macroscale cortical organization** *PNAS* **113**:12574–12579 https://doi.org/10.1073/pnas.1608282113

Matsuzawa T (2009) **Symbolic representation of number in chimpanzees** *Curr Opin Neurobiol* **19**:92–8 https://doi.org/10.1016/j.conb.2009.04.007

Matsuzawa T (1985) **Use of numbers by a chimpanzee** *Nature* **315**:57–59 https://doi.org/10.1038/315057a0

Medam T, Marzouki Y, Montant M, Fagot J (2016) **Categorization does not promote symmetry in Guinea baboons (Papio papio)** *Anim Cogn* **19**:987–998 https://doi.org/10.1007/s10071-016-1003-4

Mersad K, Kabdebon C, Dehaene-Lambertz G (2021) **Explicit access to phonetic representations in 3- month-old infants. Cognition, Special Issue in Honour of Jacques Mehler** *Cognition's founding editor* **213** https://doi.org/10.1016/j.cognition.2021.104613

Meyer T, Olson CR (2011) **Statistical learning of visual transitions in monkey inferotemporal cortex** *Proc Natl Acad Sci U S A* **108**:19401–19406 https://doi.org/10.1073/pnas.1112895108

Meyer T, Ramachandran S, Olson CR (2014) **Statistical Learning of Serial Visual Transitions by Neurons in Monkey Inferotemporal Cortex** *J Neurosci* **34**:9332–9337 https://doi.org/10.1523/JNEUROSCI.1215-14.2014

Moore MW, Durisko C, Perfetti CA, Fiez JA (2014) **Learning to Read an Alphabet of Human Faces Produces Left-lateralized Training Effects in the Fusiform Gyrus** *J Cogn Neurosci* **26**:896–913 https://doi.org/10.1162/jocn_a_00506

Murayama T, Suzuki R, Kondo Y, Koshikawa M, Katsumata H, Arai K (2017) **Spontaneous establishing of cross-modal stimulus equivalence in a beluga whale** *Sci Rep* **7** https://doi.org/10.1038/s41598-017-09925-4

Neubauer S, Hublin JJ, Gunz P (2018) **The evolution of modern human brain shape** *Sci Adv* **4** https://doi.org/10.1126/sciadv.aao5961

Neubert F-X, Mars RB, Thomas AG, Sallet J, Rushworth MFS (2014) **Comparison of human ventral frontal cortex areas for cognitive control and language with areas in monkey frontal cortex** *Neuron* **81**:700–713 https://doi.org/10.1016/j.neuron.2013.11.012

Newsome WT, Stein-Aviles JA (1999) **Nonhuman Primate Models of Visually Based Cognition** *ILAR Journal* **40**:78–91 https://doi.org/10.1093/ilar.40.2.78

Nieder A (2019) **A Brain for Numbers: The Biology of the Number Instinct**

Nieder A (2009) **Prefrontal cortex and the evolution of symbolic reference** *CurrOpinNeurobiol* **19**:99–108

Norman-Haignere S, Kanwisher NG, McDermott JH (2015) **Distinct Cortical Pathways for Music and Speech Revealed by Hypothesis-Free Voxel Decomposition** *Neuron* **88**:1281–1296 https://doi.org/10.1016/j.neuron.2015.11.035

Ogawa A, Yamazaki Y, Ueno K, Cheng K, Iriki A (2010) **Neural correlates of species-typical illogical cognitive bias in human inference** *J Cogn Neurosci* **22**:2120–2130 https://doi.org/10.1162/jocn.2009.21330

Pallier C, Devauchelle A-D, Dehaene S (2011) **Cortical representation of the constituent structure of sentences** *PNAS* **108**:2522–2527 https://doi.org/10.1073/pnas.1018711108

Palomero-Gallagher N, Zilles K (2019) **Differences in cytoarchitecture of Broca's region between human, ape and macaque brains. Cortex** *The Evolution of the Mind and the Brain* **118**:132–153 https://doi.org/10.1016/j.cortex.2018.09.008

Patel AD (2010) **Music, Language, and the Brain**

Penn DC, Holyoak KJ, Povinelli DJ (2008) **Darwin's mistake: explaining the discontinuity between human and nonhuman minds** *Behav Brain Sci* **31**:109–30 https://doi.org/10.1017/S0140525X08003543

Pepperberg IM (2009) **The Alex studies: cognitive and communicative abilities of grey parrots**

Pepperberg IM, Carey S (2012) **Grey parrot number acquisition: the inference of cardinal value from ordinal position on the numeral list** *Cognition* **125**:219–232 https://doi.org/10.1016/j.cognition.2012.07.003

Perszyk DR, Waxman SR (2019) **Infants' advances in speech perception shape their earliest links between language and cognition** *Sci Rep* **9** https://doi.org/10.1038/s41598-019-39511-9

Petkov CI, Kayser C, Augath M, Logothetis NK (2009) **Optimizing the imaging of the monkey auditory cortex: sparse vs. continuous fMRI** *Magn ResonImaging* **27**:1065–1073

Petrides M, Tomaiuolo F, Yeterian EH, Pandya DN (2012) **The prefrontal cortex: comparative architectonic organization in the human and the macaque monkey brains** *Cortex* **48**:46–57

Piazza M (2010) **Neurocognitive start-up tools for symbolic number representations. Trends in Cognitive Sciences, Special Issue: Space** *Time and Number* **14**:542–551 https://doi.org/10.1016/j.tics.2010.09.008

Pinel P, Thirion B, Meriaux S, Jobert A, Serres J, Le Bihan D, Poline JB, Dehaene S. (2007) **Fast reproducible identification and large-scale databasing of individual functional cognitive networks** *BMC neuroscience* **8**

Premack D (1971) **Language in chimpanzee** *Science* **172**:808–822

Richter D, Ekman M, Lange FP de (2018) **Suppressed Sensory Response to Predictable Object Stimuli throughout the Ventral Visual Stream** *J Neurosci* **38**:7452–7461 https://doi.org/10.1523/JNEUROSCI.3421-17.2018

Rilling JK (2014) **Comparative primate neuroimaging: insights into human brain evolution** *Trends in Cognitive Sciences* **18**:46–55 https://doi.org/10.1016/j.tics.2013.09.013

Rilling JK, Glasser MF, Jbabdi S, Andersson J, Preuss TM (2012) **Continuity, Divergence, and the Evolution of Brain Language Pathways** *Front Evol Neurosci* **3** https://doi.org/10.3389/fnevo.2011.00011

Rilling JK, Glasser MF, Preuss TM, Ma X, Zhao T, Hu X, Behrens TEJ (2008) **The evolution of the arcuate fasciculus revealed with comparative DTI** *Nat Neurosci* **11**:426–428 https://doi.org/10.1038/nn2072

Roelfsema PR, Treue S (2014) **Basic neuroscience research with nonhuman primates: a small but indispensable component of biomedical research** *Neuron* **82**:1200–1204

Sablé-Meyer M, Fagot J, Caparos S, van Kerkoerle T, Amalric M, Dehaene S. (2021) **Sensitivity to geometric shape regularity in humans and baboons: A putative signature of human singularity** *ProcNatlAcadSciUSA* **118**

Schenker NM, Hopkins WD, Spocter MA, Garrison AR, Stimpson CD, Erwin JM, Hof PR, Sherwood CC (2010) **Broca's area homologue in chimpanzees (Pan troglodytes): probabilistic mapping, asymmetry, and comparison to humans** *Cereb Cortex* **20**:730–742 https://doi.org/10.1093/cercor/bhp138

Schusterman R, Kastak D (1998) **Functional equivalence in a California sea lion: relevance to animal social and communicative interactions** *Anim Behav* **55**:1087–1095 https://doi.org/10.1006/anbe.1997.0654

Scott BH, Mishkin M (2016) **Auditory short-term memory in the primate auditory cortex** *Brain Res* **1640**:264–277 https://doi.org/10.1016/j.brainres.2015.10.048

Sestieri C, Shulman GL, Corbetta M (2010) **Attention to Memory and the Environment: Functional Specialization and Dynamic Competition in Human Posterior Parietal Cortex** *J Neurosci* **30**:8445–8456 https://doi.org/10.1523/JNEUROSCI.4719-09.2010

Shibata M, Pattabiraman K, Lorente-Galdos B, Andrijevic D, Kim S-K, Kaur N, Muchnik SK, Xing X, Santpere G, Sousa AMM, Sestan N (2021) **Regulation of prefrontal patterning and connectivity by retinoic acid** *Nature* **598**:483–488 https://doi.org/10.1038/s41586-021-03953-x

Shum J, Hermes D, Foster BL, Dastjerdi M, Rangarajan V, Winawer J, Miller KJ, Parvizi J (2013) **A brain area for visual numerals** *J Neurosci* **33**:6709–15 https://doi.org/10.1523/JNEUROSCI.4558-12.2013

Sidman M, Rauzin R, Lazar R, Cunningham S, Tailby W, Carrigan P (1982) **A search for symmetry in the conditional discriminations of rhesus monkeys, baboons, and children** *J Exp Anal Behav* **37**:23–44 https://doi.org/10.1901/jeab.1982.37-23

Smaers JB, Gómez-Robles A, Parks AN, Sherwood CC (2017) **Exceptional Evolutionary Expansion of Prefrontal Cortex in Great Apes and Humans** *Current Biology* **27**:714–720 https://doi.org/10.1016/j.cub.2017.01.020

Song Y, Tian M, Liu J (2012) **Top-down processing of symbolic meanings modulates the visual word form area** *J Neurosci* **32**:12277–83 https://doi.org/10.1523/JNEUROSCI.1874-12.2012

Spelke E, Gentner D, Goldin- Meadow S (2003) **What makes us smart? Core knowledge and natural language** *Language in Mind*

Srihasam K, Mandeville JB, Morocz IA, Sullivan KJ, Livingstone MS (2012) **Behavioral and Anatomical Consequences of Early versus Late Symbol Training in Macaques** *Neuron* **73**:608–619 https://doi.org/10.1016/j.neuron.2011.12.022

Takemura H, Pestilli F, Weiner KS, Keliris GA, Landi SM, Sliwa J, Ye FQ, Barnett MA, Leopold DA, Freiwald WA, Logothetis NK, Wandell BA (2017) **Occipital White Matter Tracts in Human and Macaque** *Cereb Cortex* **27**:3346–3359 https://doi.org/10.1093/cercor/bhx070

Tasserie J, Grigis A, Uhrig L, Dupont M, Amadon A, Jarraya B (2020) **Pypreclin: An automatic pipeline for macaque functional MRI preprocessing** *NeuroImage* **207** https://doi.org/10.1016/j.neuroimage.2019.116353

Thiebaut de Schotten M, Dell'Acqua F, Valabregue R, Catani M (2012) **Monkey to human comparative anatomy of the frontal lobe association tracts** *Cortex* **48**:82–96 https://doi.org/10.1016/j.cortex.2011.10.001

Uhrig L, Dehaene S, Jarraya B. (2014) **A hierarchy of responses to auditory regularities in the macaque brain** *JNeurosci* **34**:1127–1132

Vanduffel W, Fize D, Mandeville JB, Nelissen K, Van Hecke P, Rosen BR, Tootell RBH, Orban GA. (2001) **Visual motion processing investigated using contrast agent-enhanced fMRI in awake behaving monkeys** *Neuron* **32**:565–577

Vanduffel W, Fize D, Peuskens H, Denys K, Sunaert S, Todd JT, Orban GA (2002) **Extracting 3D from Motion: Differences in Human and Monkey Intraparietal Cortex** *Science* **298**:413–415 https://doi.org/10.1126/science.1073574

Vergnieux V, Vogels R (2020) **Statistical Learning Signals for Complex Visual Images in Macaque Early Visual Cortex** *Frontiers in Neuroscience* **14**

Wang L, Amalric M, Fang W, Jiang X, Pallier C, Figueira S, Sigman M, Dehaene S (2019) **Representation of spatial sequences using nested rules in human prefrontal cortex** *NeuroImage* **186**:245–255 https://doi.org/10.1016/j.neuroimage.2018.10.061

Wang L, Uhrig L, Jarraya B, Dehaene S (2015) **Representation of Numerical and Sequential Patterns in Macaque and Human Brains** *Curr Biol* **25**:1966–1974 https://doi.org/10.1016/j.cub.2015.06.035

Warren JM (1974) **Possibly unique characteristics of learning by Primates** *Journal of Human Evolution* **3**:445–454 https://doi.org/10.1016/0047-2484(74)90004-9

Wikman P, Rinne T, Petkov CI (2019) **Reward cues readily direct monkeys' auditory performance resulting in broad auditory cortex modulation and interaction with sites along cholinergic and dopaminergic pathways** *Sci Rep* **9** https://doi.org/10.1038/s41598-019-38833-y

Wilson B, Marslen-Wilson WD, Petkov CI (2017) **Conserved Sequence Processing in Primate Frontal Cortex** *Trends Neurosci* **40**:72–82 https://doi.org/10.1016/j.tins.2016.11.004

Wise SP (2008) **Forward frontal fields: phylogeny and fundamental function** *Trends Neurosci* **31**:599–608 https://doi.org/10.1016/j.tins.2008.08.008

Woods DL, Herron TJ, Cate AD, Yund EW, Stecker GC, Rinne T, Kang X (2010) **Functional properties of human auditory cortical fields** *Front Syst Neurosci* **4** https://doi.org/10.3389/fnsys.2010.00155

Xu F, Cote M, Baker A (2005) **Labeling guides object individuation in 12-month-old infants** *Psychol Sci* **16**:372–7 https://doi.org/10.1111/j.0956-7976.2005.01543.x

Yang C. (2013) **Ontogeny and phylogeny of language** *PNAS* **201216803** https://doi.org/10.1073/pnas.1216803110

Zhang H, Zhen Y, Yu S, Long T, Zhang B, Jiang X, Li J, Fang W, Sigman M, Dehaene S, Wang L (2022) **Working Memory for Spatial Sequences: Developmental and Evolutionary Factors in Encoding Ordinal and Relational Structures** *J Neurosci* **42**:850–864 https://doi.org/10.1523/JNEUROSCI.0603-21.2021

## Author information

### Timo van Kerkoerle

Cognitive Neuroimaging Unit, CEA, INSERM, Université Paris-Saclay, NeuroSpin center, 91191 Gif/Yvette, France
**For correspondence:** timo@neuroscience.visio
ORCID iD: 0000-0003-1935-8216

### Louise Pape

Cognitive Neuroimaging Unit, CEA, INSERM, Université Paris-Saclay, NeuroSpin center, 91191 Gif/Yvette, France

**Milad Ekramnia**
Cognitive Neuroimaging Unit, CEA, INSERM, Université Paris-Saclay, NeuroSpin center, 91191 Gif/Yvette, France
ORCID iD: 0000-0003-4031-3055

**Xiaoxia Feng**
Cognitive Neuroimaging Unit, CEA, INSERM, Université Paris-Saclay, NeuroSpin center, 91191 Gif/Yvette, France

**Jordy Tasserie**
Cognitive Neuroimaging Unit, CEA, INSERM, Université Paris-Saclay, NeuroSpin center, 91191 Gif/Yvette, France, Center for Brain Circuit Therapeutics Department of Neurology Brigham & Women's Hospital, Harvard Medical School, Boston, MA, USA
ORCID iD: 0000-0001-9626-7823

**Morgan Dupont**
Cognitive Neuroimaging Unit, CEA, INSERM, Université Paris-Saclay, NeuroSpin center, 91191 Gif/Yvette, France

**Xiaolian Li**
Department of Neurosciences, Laboratory of Neuro- and Psychophysiology, KU Leuven Medical School, Leuven 3000, Belgium, Leuven Brain Institute, KU Leuven, Leuven 3000, Belgium

**Bechir Jarraya**
Cognitive Neuroimaging Unit, CEA, INSERM, Université Paris-Saclay, NeuroSpin center, 91191 Gif/Yvette, France, Leuven Brain Institute, KU Leuven, Leuven 3000, Belgium

**Wim Vanduffel**
Department of Neurosciences, Laboratory of Neuro- and Psychophysiology, KU Leuven Medical School, Leuven 3000, Belgium, Leuven Brain Institute, KU Leuven, Leuven 3000, Belgium, Athinoula A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Charlestown, MA 02129, USA, Department of Radiology, Harvard Medical School, Boston, MA 02144, USA
ORCID iD: 0000-0002-9399-343X

**Stanislas Dehaene**
Cognitive Neuroimaging Unit, CEA, INSERM, Université Paris-Saclay, NeuroSpin center, 91191 Gif/Yvette, France, Collège de France, Université Paris-Sciences-Lettres (PSL), 11 Place Marcelin Berthelot, 75005 Paris, France
ORCID iD: 0000-0002-7418-8275

**Ghislaine Dehaene-Lambertz**
Cognitive Neuroimaging Unit, CEA, INSERM, Université Paris-Saclay, NeuroSpin center, 91191 Gif/Yvette, France
**For correspondence:** ghislaine.dehaene@cea.fr
ORCID iD: 0000-0003-2221-9081

## Editors

Reviewing Editor
**Andrea Martin**
Max Planck Institute for Psycholinguistics, Netherlands

Senior Editor
**Timothy Behrens**
University of Oxford, United Kingdom

## Reviewer #1 (Public Review):

Kerkoerle and colleagues present a very interesting comparative fMRI study in humans and monkeys, assessing neural responses to surprise reactions at the reversal of a previously learned association. The implicit nature of this task, assessing how this information is represented without requiring explicit decision-making, is an elegant design. The paper reports that both humans and monkeys show neural responses across a range of areas when presented with incongruous stimulus pairs. Monkeys also show a surprise response when the stimuli are presented in a reversed direction. However, humans show no such surprise response based on this reversal, suggesting that they encode the relationship reversibly and bidirectionally, unlike the monkeys. This has been suggested as a hallmark of symbolic representation, that might be absent in nonhuman animals.

I find this experiment and the results quite compelling, and the data do support the hypothesis that humans are somewhat unique in their tendency to form reversible, symbolic associations. I think that an important strength of the results is that the critical finding is the presence of an interaction between congruity and canonicity in macaques, which does not appear in humans. These results go a long way to allay concerns I have about the comparison of many human participants to a very small number of macaques.

I understand the impossibility of testing 30+ macaques in an fMRI experiment. However, I think it is important to note that differences necessarily arise in the analysis of such datasets. The authors report that they use '...identical training, stimuli, and whole-brain fMRI measures'. However, the monkeys (in experiment 1) actually required 10 times more training. More importantly, while the fMRI measures are the same, group analysis over 30+ individuals is inherently different from comparing only 2 macaques (including smoothing and averaging away individual differences that might be more present in the monkeys, due to the much smaller sample size).

Despite this, the results do appear to show that macaques show the predicted interaction effect (even despite the sample size), while humans do not. I think this is quite convincing, although had the results turned out differently (for example an effect in humans that was absent in macaques), I think this difference in sample size would be considerably more concerning.

I would also note that while I agree with the authors' conclusions, it is notable to me that the congruity effect observed in humans (red vs blue lines in Fig. 2B) appears to be far more pronounced than any effect observed in the macaques (Fig. 3C-3). Again, this does not challenge the core finding of this paper but does suggest methodological or possibly motivational/attentional differences between the humans and the monkeys (or, for example, that the monkeys had learned the associations less strongly and clearly than the humans).

This is a strong paper with elegant methods and makes a worthwhile contribution to our understanding of the neural systems supporting symbolic representations in humans, as opposed to other animals.

## Reviewer #2 (Public Review):

In their article titled "Brain mechanisms of reversible symbolic reference: a potential singularity of the human brain", van Kerkoerle et al address the timely question of whether non-human primates (rhesus macaques) possess the ability for reverse symbolic inference as observed in humans. Through an fMRI experiment in both humans and monkeys, they analyzed the bold signal in both species while observing audio-visual and visual-visual stimuli pairs that had been previously learned in a particular direction. Remarkably, the findings pertaining to humans revealed that a broad brain network exhibited increased activity in response to surprises occurring in both the learned and reverse directions. Conversely, in monkeys, the study uncovered that the brain activity within sensory areas only responded to the learned direction but failed to exhibit any discernible response to the reverse direction. These compelling results indicate that the capacity for reversible symbolic inference may be unique to humans.

In general, the manuscript is skillfully crafted and highly accessible to readers. The experimental design exhibits originality, and the analyses are tailored to effectively address the central question at hand. Although the first experiment raised a number of methodological inquiries, the subsequent second experiment thoroughly addresses these concerns and effectively replicates the initial findings, thereby significantly strengthening the overall study. Overall, this article is already of high quality and brings new insight into human cognition.

I identified three weaknesses in the manuscript:
- One major issue in the study is the absence of significant results in monkeys. Indeed, authors draw conclusions regarding the lack of significant difference in activity related to surprise in the multi-demand network (MDN) in the reverse congruent versus reverse incongruent conditions. Although the results are convincing (especially with the significant interaction between congruency and canonicity), the article could be improved by including additional analyses in a priori ROI for the MDN in monkeys (as well as in humans, for comparison).
- While the authors acknowledge in the discussion that the number of monkeys included in the study is considerably lower compared to humans, it would be informative to know the variability of the results among human participants.
- Some details are missing in the methods.

## Reviewer #3 (Public Review):

This study investigates the hypothesis that humans (but not non-human primates) spontaneously learn reversible temporal associations (i.e., learning a B-A association after only being exposed to A-B sequences), which the authors consider to be a foundational property of symbolic cognition. To do so, they expose humans and macaques to 2-item sequences (in a visual-auditory experiment, pairs of images and spoken nonwords, and in a visual-visual experiment, pairs of images and abstract geometric shapes) in a fixed temporal order, then measure the brain response during a test phase to congruent vs. incongruent pairs (relative to the trained associations) in canonical vs. reversed order (relative to the presentation order used in training). The advantage of neuroimaging for this question is that it removes the need for a behavioral test, which non-human primates can fail for reasons unrelated to the cognitive construct being investigated. In humans, the researchers find statistically indistinguishable incongruity effects in both directions (supporting a spontaneous reversible association), whereas in monkeys they only find incongruity effects in the canonical direction (supporting an association but a lack of spontaneous reversal). Although the precise pattern of activation varies by experiment type (visual-auditory vs. visual-visual) in both species, the authors point out that some of the regions involved are also those that are most anatomically different between humans and other primates. The authors

interpret their finding to support the hypothesis that reversible associations, and by extension symbolic cognition, is uniquely human.

This study is a valuable complement to prior behavioral work on this question. However, I have some concerns about methods and framing.

Methods - Design issues:

1. The authors originally planned to use the same training/testing protocol for both species but the monkeys did not learn anything, so they dramatically increased the amount of training and evaluation. By my calculation from the methods section, humans were trained on 96 trials and tested on 176, whereas the monkeys got an additional 3,840 training trials and 1,408 testing trials. The authors are explicit that they continued training the monkeys until they got a congruity effect. On the one hand, it is commendable that they are honest about this in their write-up, given that this detail could easily be framed as deliberate after the fact. On the other hand, it is still a form of p-hacking, given that it's critical for their result that the monkeys learn the canonical association (otherwise, the critical comparison to the non-canonical association is meaningless).

2. Between-species comparisons are challenging. In addition to having differences in their DNA, human participants have spent many years living in a very different culture than that of NHPs, including years of formal education. As a result, attributing the observed differences to biology is challenging. One approach that has been adopted in some past studies is to examine either young children or adults from cultures that don't have formal educational structures. This is not the approach the authors take. This major confound needs to minimally be explicitly acknowledged up front.

3. Humans have big advantages in processing and discriminating spoken stimuli and associating them with visual stimuli (after all, this is what words are in spoken human languages). Experiment 2 ameliorates these concerns to some degree, but still, it is difficult to attribute the failure of NHPs to show reversible associations in Experiment 1 to cognitive differences rather than the relative importance of sound string to meaning associations in the human vs. NHP experiences.

4. More minor: The localizer task (math sentences vs. other sentences) makes sense for math but seems to make less sense for language: why would a language region respond more to sentences that don't describe math vs. ones that do?

Methods - Analysis issues:

5. The analyses appear to "double dip" by using the same data to define the clusters and to statistically test the average cluster activation (Kriegeskorte et al., 2009). The resulting effect sizes are therefore likely inflated, and the p-values are anticonservative.

Framing:

6. The framing ("Brain mechanisms of reversible symbolic reference: A potential singularity of the human brain") is bigger than the finding (monkeys don't spontaneously reverse a temporal association but humans do). The title and discussion are full of buzzy terms ("brain mechanisms", "symbolic", and "singularity") that are only connected to the experiments by a debatable chain of assumptions.

First, this study shows relatively little about brain "mechanisms" of reversible symbolic associations, which implies insights into how these associations are learned, recognized, and represented. But we're only given standard fMRI analyses that are quite inconsistent across similar experimental paradigms, with purely suggestive connections between these spatial patterns and prior work on comparative brain anatomy.

Second, it's not clear what the relationship is between symbolic cognition and a propensity to spontaneously reverse a temporal association. Certainly, if there are inter-species differences in learning preferences this is important to know about, but why is this construed as a difference in the presence or absence of symbols? Because the associations aren't used in any downstream computation, there is not even any way for participants to know which is the sign and which is the signified: these are merely labels imposed by the researchers on a sequential task.

Third, the word "singularity" is both problematically ambiguous and not well supported by the results. "Singularity" is a highly loaded word that the authors are simply using to mean "that which is uniquely human". Rather than picking a term with diverse technical meanings across fields and then trying to restrict the definition, it would be better to use a different term. Furthermore, even under the stated definition, this study performed a single pairwise comparison between humans and one other species (macaques), so it is a stretch to then conclude (or insinuate) that the "singularity" has been found (see also pt. 2 above).

7. Related to pt. 6, there is circularity in the framing whereby the authors say they are setting out to find out what is uniquely human, hypothesizing that the uniquely human thing is symbols, and then selecting a defining trait of symbols (spontaneous reversible association) *because* it seems to be uniquely human (see e.g., "Several studies previously found behavioral evidence for a uniquely human ability to spontaneously reverse a learned association (Imai et al., 2021; Kojima, 1984; Lipkens et al., 1988; Medam et al., 2016; Sidman et al., 1982), and such reversibility was therefore proposed as a defining feature of symbol representation reference (Deacon, 1998; Kabdebon and Dehaene-Lambertz, 2019; Nieder, 2009).", line 335). They can't have it both ways. Either "symbol" is an independently motivated construct whose presence can be independently tested in humans and other species, or it is by fiat synonymous with the "singularity". This circularity can be broken by a more modest framing that focuses on the core research question (e.g., "What is uniquely human? One possibility is spontaneous reversal of temporal associations.") and then connects (speculatively) to the bigger conceptual landscape in the discussion ("Spontaneous reversal of temporal associations may be a core ability underlying the acquisition of mental symbols").