Hierarchical surprise signals in naturalistic violation of expectations

Vincent Plikat[1,2,3,*,+], Pablo R. Grassi[1,2,3,*,+], Julius Frack[4], Andreas Bartels[1,2,3,+]
[1] Department of Psychology, University of Tübingen, Tübingen, Germany
[2] Centre for Integrative Neuroscience, Tübingen, Germany
[3] Max-Planck Institute for Biological Cybernetics, Tübingen, Germany
[4] Rocket Magic, Tübingen, Germany
*, equal contribution
+, corresponding authors

Abstract

Surprise responses signal both, high-level cognitive alerts that information is missing, and increasingly specific back-propagating error signals that allow updates in processing nodes. Studying surprise is hence central for cognitive neuroscience to understand internal world representations and learning. Yet, only few prior studies used naturalistic stimuli targeting our high-level understanding of the world. Here, we use magic tricks in an fMRI experiment to investigate neural responses to violations of core assumptions held by humans about the world. We showed participants naturalistic videos of three types of magic tricks, involving objects appearing, changing color, or disappearing, along with control videos without any violation of expectation. Importantly, the same videos were presented with and without prior knowledge about the tricks' explanation. Results revealed generic responses in frontoparietal areas, together with responses specific to each of the three trick types in posterior sensory areas. A subset of these regions, the midline areas of the default mode network (DMN), showed surprise activity that depended on prior knowledge. Equally, sensory regions showed sensitivity to prior knowledge, reflected in differing decoding accuracies. These results suggest a hierarchy of surprise signals involving three processing stages: first, generic processing of violation of expectations in frontoparietal areas. Second, surprise signals in sensory regions that are specific to the processed feature, demonstrating propagation of surprise signals to mid-level visual areas. Third, prior-knowledge dependent processing of surprise signals in parts of the DMN, showing its involvement in internal state monitoring and in making sense of events as they unfold in time.

Keywords: surprise, magic, predictive coding, violation of expectation, intuitive physics, fMRI

Word count: Abstract: 250, **Introduction:** 900, **Discussion:** 1818, **Total:** ca. 13000 (w/o refs)

Introduction

Prior experience and "intuitive" knowledge about the physical world guide our perception and allow for a meaningful interaction with the environment. They set up constraints on our expectations based on what we believe to be possible in the world. For example, prior world-knowledge informs us that objects do not vanish of existence if occluded (object permanency), that objects tend to keep their features (feature constancy), that objects cannot pass through other objects (solidity), that objects to not appear out of the blue, and so forth. Informed expectations based on these intuitive physical priors allow us to quickly make sense of incoming sensory information. Such expectations have been shown to strongly modulate perception (de Lange et al., 2018). For example, a "light-from-above" prior constrains depth perception from shading (Adams et al., 2004), and knowledge that objects cannot occupy the same place at the same time explains our inability to perceive two objects simultaneously in bistable perception (Hohwy et al., 2008).

Accordingly, perception can be understood as an inferential process in which top-down informed expectations are matched with incoming sensory information (Friston, 2010, 2005; Lee and Mumford, 2003; Rao and Ballard, 1999). In this "predictive processing" framework deviations between expectations and incoming data (i.e., prediction errors) are used to update an internal model of the world at different levels of complexity and abstraction. Higher-level priors (like object permanency) represent more abstract aspects of the world and constrain lower-level inferences that represent more immediate features of the world (like a particular object) (Clark, 2013; Hohwy, 2014). In this context, intuitive physical knowledge can be thought of as an internal model representing different aspects of the causal structure of the physical world, not unsimilar to a "physics engine" in a virtual environment (Battaglia et al., 2013).

This prior knowledge of physical principles can be studied using events that seemingly violate them. For example, evidence from developmental psychology using so-called violation of expectation (VOE) paradigms suggest that infants acquire important aspects about the workings of the physical world during the first year of life, such as object permanency and solidity (Hespos et al., 2009; Wang, 2004; Wynn, 1992). However, it is largely unclear how surprise-responses relating to intuitive physical principles are represented in the human brain. Most previous neuroimaging studies investigated responses to lower-level VOE, e.g. using paradigms involving infrequent (and thus unexpected) stimuli (e.g., Egner et al., 2010; Kok et al., 2012a; Todorovic et al., 2011; Wessel et al., 2012), or omission of expected stimuli (Kok et al., 2012b; SanMiguel et al., 2013; e.g., Wacongne et al., 2011). These studies revealed lower-level stimulus-specific prediction errors in modality- and feature-specific areas. In contrast, studies investigating VOE of higher-level physical principles using more complex stimuli such as computer generated animations (Bardi et al., 2017; Liu et al., 2024) or naturalistic videos showing magic tricks (Danek et al., 2015; Parris et al., 2009) revealed higher-level surprise signals in frontoparietal areas. This is in line with a frontoparietal role in the representation of physical concepts (Fischer et al., 2016;

Schwettmann et al., 2019). However, higher-level VOE studies have hitherto not observed VOE-related activity in sensory areas.

Here, we set out to unify and resolve this longstanding divergence of results between high-level and low-level VOE paradigms by using a novel VOE paradigm. Our paradigm was designed to uncover hierarchical VOE signals of the brain's internal world model. It contained a battery of standardized magic videos that we presented to human participants while measuring fMRI responses to investigate: 1) which regions are generally involved when viewing natural videos that violate physical principles, 2) whether specific types of violations, like appearance of objects, or changes of color, modulate sensory areas known to process the feature concerned, 3) whether knowledge of the explanation of a given magic trick modulates the observed VOE activity.

To this aim, we created and validated videos for a naturalistic VOE paradigm showing either dedicated magic tricks (to create the illusion of seemingly impossible events to actually occur, cf. Grassi and Bartels, 2021) or matched control actions that involved no violation of physical principles. The VOE videos were designed to evoke surprise responses related to unexpected object appearance, disappearance of objects, and feature change (color-changing objects). They were performed by a professional magician (Julius Frack), and each trick-type was performed using three common objects (balls, playing cards and pencils). Each trick was presented before and after revealing the method of the tricks. This allowed us to compare responses with and without VOE using identical videos.

Univariate analyses revealed a hierarchy of surprise signals: frontoparietal areas, including areas of the default mode network (DMN), were involved when perceiving events violating physical principles regardless of the type of trick used. In contrast, posterior sensory areas were modulated specifically by the type of expectation-violation, such as color-processing medial fusiform cortex by color change, and object selective LOC by the appearance of objects. Controls indicate that their modulation is due to the feature-specific surprise and not due to the feature-change. Multivariate analyses extended the results: information about the specific types of expectation-violations was exclusively encoded in posterior regions, and significantly decodable down to the earliest levels of cortical visual processing (V1-V3). Additionally, decoding accuracy significantly decreased with prior knowledge, revealing a reduction in surprise signals. Together, our results demonstrate a generic response in frontoparietal areas to violation of physical principles, along with concurrent representations of specific expected information in early sensory areas.

Methods

Participants

We performed fMRI on 27 subjects. Three subjects were excluded from data analysis due to excessive movement and/or sleepiness during the scanning sessions. Data from a total of 24 subjects were analyzed (16 female; 8 male; mean age 24.4 ± 4.3 SD years). All subjects had normal or corrected to normal vision, no history of neurological impairments nor contraindication for fMRI. Participants had no expertise as magicians and were naive to the magic tricks used. Participants provided written informed consent prior to the experiment. The study was approved by the ethics committee of the University Clinic Tübingen and was conducted in accord with the Declaration of Helsinki.

VOE stimuli

Video recordings. A set of 63 videos were created for the VOE study. In accord with previous work (Parris et al., 2009), we created videos for three different conditions: the videos showed either magic tricks (magic condition, 18 videos), similar actions without a magic event (control condition, 18 videos) or unusual actions with the objects used in the magic tricks (unusual condition, 9 videos). We included the unusual condition to investigate neural correlates of surprise in a similar setting while not violating any physical concept (cf. Parris et al., 2009). We further created explanation videos showing how each of the magic tricks was achieved (18 videos).

All videos were performed by professional illusionist Julius Frack in a standardized setting consisting of a black background and a black table (see Figure 1A for an example). To investigate the effect of specific VOE, we presented magic tricks showing three different violations of physical principles: appearances (A) (i.e., a red object appears), color changes (C) (i.e., a red object changes color to blue) and vanishes (V) (i.e., a red object disappears) (see Figure 1C). To generalize across objects, we used three easily distinguishable objects: balls, playing cards and thick pencils. Each object was used equally often in each trick type. Finally, we created two versions of each type of VOE for each object (e.g., two different videos showing a red ball changing to a blue ball using different methods).

Full set of stimuli. Each of these magic tricks had a matched control video that showed the same sequences of actions as the trick, but without VOE (e.g., following the same actions a ball would not change its color). Thus, we had a total of 18 magic videos (3 type of VOEs × 3 objects × 2 methods), with 18 matching control videos and nine unusual videos (3 unusual actions per used object). As the solutions to the magic tricks were revealed in distinct sets throughout the experiment, we used a variety of different methods for each type of VOE. This ensured that participants were only able to infer trick solutions they were intended to understand.

Luminance and durations. Tricks were recorded in a standardized setting under the same lighting conditions. To balance out remaining inequalities, custom MATLAB (MathWorks, Natick, MA) scripts were used to standardize the videos (resolution of 1920 x 1080 with 25 frames per second), which were filmed on different days and had different lengths. We manually applied white-balance using 5 to 10 selected white-pixels for all videos of a same day, matched the luminance and contrast of the videos based their first frame, and shortened the videos to be no longer than 14 seconds. The final duration of the videos was 12.8 s ± 1.08 s (mean ± SD).

Stimulus presentation. Stimuli were presented using MATLAB 2019b using Psychtoolbox3 (version 3.0.16 http://psychtoolbox.org/) on a Linux computer and back-projected to a translucent screen mounted at the rear of the scanner bore using a VPIXX Pro projector (VPixx Technologies, Saint-Bruno-de-Montarville, Canada) at a frame rate of 144 Hz. Participants viewed the screen (26.1 x 14.7 visual degrees) via a mirror mounted on the 64-channel head coil (Siemens, Erlangen, Germany) at a distance of 105 cm. To center the stimulus presentation, we cropped 160 pixels from the left and right side of the videos that only showed a black background. Accordingly, the shown part of the videos covered 21.8 x 14.7 visual degrees (1600 x 1080 pixels).

Behavioral evaluation of stimuli. Prior to the fMRI experiment we performed two psychophysics experiments with a total of 18 subjects (nine subjects in each experiment) to ensure the suitability of our stimuli and to select the magic tricks to be used in the fMRI experiment (see results in Supplementary Section *Behavioral evaluation of stimuli*). Participants were shown the videos using the same design as in the fMRI experiment (see Figure 1B). Participants were surprised when viewing the magic videos, even though they knew they were observing tricks (Grassi et al., 2024) and systematically reported to be more surprised when viewing the magic videos compared to the matched controls. Moreover, surprise responses to the VOE shown in the magic videos were reduced after providing an explanation to the tricks.

Experimental design and procedure

fMRI runs. Each run consisted of a total of 24 video trials: two repetitions of the six unique magic videos of one object (three types of VOE, each type recorded in two versions), resulting in 12 magic trick presentations; the six matching control videos (shown once); and two repetitions of three additional unusual videos, resulting in six unusual video presentations (see Figure 1D). The resulting 24 video trials were presented in a pseudo-random order that avoided repetitions of identical videos (different randomizations across runs).

Paradigm. The fMRI experiment consisted of three sets with four experimental fMRI runs each. Each set presented videos of only one object (i.e., balls, playing cards or pencils). The

presentation order of the object sets was counterbalanced across subjects. After the first two fMRI runs in a set (pre-revelation runs) the method behind each magic trick in the set was revealed by showing the participants each of the magic tricks again together with the matching revelation video. Subjects could watch the videos as often as they wished and were asked to confirm per button press that they understood how each of the tricks was achieved. Thereafter, two more runs showing the same videos were performed (post-revelation runs). Together, each set consisted of two fMRI runs before and two fMRI runs after the explanation of the tricks. A visualization of the experimental design is shown in Figure 1E. We hypothesized that providing the explanation of the tricks would decrease VOE responses because participants would adjust their expectations (e.g., knowing where the seemingly disappearing objects were concealed). Consequently, our experimental design allowed us to compare responses with and without VOE using identical stimuli.

Individual trials. Each video trial was presented for 14 seconds. In case a video was shorter (e.g., 13 seconds), the last frame of the video was shown for the remaining time (e.g., for 1 second). Participants were informed about this. Moreover, the first 500 ms of each video showed a central cue (1.3 visual degrees) indicating whether the video was going to be a magic (M) or a non-magic (X) video (i.e., control or unusual actions). After each video presentation participants were asked to rate from 1 to 5 how surprising the content of the video was (1 = not surprising, 5 = very surprising). They had two seconds to respond, after which the next trial started (total trial duration = 16 seconds, see Figure 1B). Behavioral surprise ratings were given by the subjects using a button-box with five keys.

Trial cues. We included the cue (M or X) at the beginning of the trial to prevent participants from being surprised by contextual or serial effects. For example, since we tried to create control and magic videos that are visually as similar as possible, one could easily mistake a control video for a magic trick at the beginning of the video. The missing magic trick could then be surprising for participants expecting one. Yet, this was not the kind of surprise we wanted to investigate. We hypothesized that prior knowledge about the content of a trial (magic vs. no-magic) should prevent this from happening. Moreover, to reduce predictability of video content we flipped each video on every second presentation of the same video horizontally. For example, if the magician performed a color changing card trick on his left-hand side in the first presentation of the video, the next presentation of the same video showed the color changing card trick on his right-hand side.

Pre-scan instructions. Before scanning, we instructed subjects about the task and design of the experiment (i.e., the set design, meaning of trial cues, explanation videos, etc.). Participants were informed that videos would show either magic tricks, control actions or unusual actions. Moreover, participants were shown a magic and matching control example video (not used in the actual

experiment) to give them an impression about the kind and duration of the videos they were about to see. Further, to prevent participants from watching the videos in a “problem-solving” attitude, we instructed them to passively watch and enjoy the videos without trying to get behind the method of the tricks, as these would be explained during the experiment.

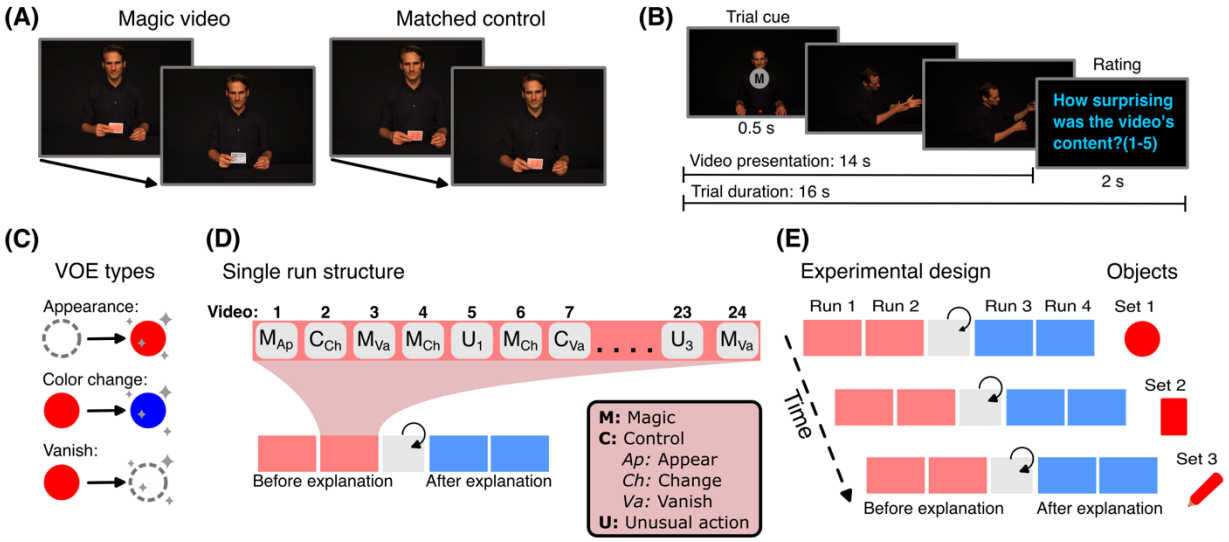


Figure 1. A, example of a color-changing card magic trick (left) and its corresponding matched control video (right). **B**, shown is the timeline of a single trial. Every video started with a central cue lasting 500 ms indicating whether the video will show a magic trick (M) or not (X). Each video presentation lasted 14 seconds, if a video happened to be shorter than 14 s, the last frame of the video was shown until the total duration was 14 s as a static image. After the video, subjects had 2 seconds to answer how surprising the video’s content was. Note that the text was displayed in one line and did not cover the whole screen. **C**, shown are the three different types of VOE used (i.e., "magic events"). We created videos showing magic tricks using three objects (balls, playing cards and pencils) that showed either an unexpected object appear (Ap), change color (from red to blue) (Ch) or vanish (Va). For each object-VOE combination, we created two tricks (i.e., we had two color-changing card tricks). **D**, schematic example of a single experimental set. One set consisted of four runs – two before and two after explanation runs (i.e., pre and post revelation). Between pre- and post-revelation runs we showed participants videos explaining the magic tricks performed during the corresponding set. Each run showed 24 videos with a pseudo-randomized order ensuring that the same video was not presented in two consecutive trials (twelve magic videos [M], six matched controlled videos [C] and six unusual actions [U]). **E**, complete experimental design showing all three sets. The experiment was divided into three sets, one for each of the objects. Each set was divided into four experimental runs, where each run showed all the videos of an object but in a randomized order. After the second fMRI run the methods behind the tricks shown in the set were revealed, dividing the runs into *before* and *after* the explanation of the tricks (pre- and post-revelation, respectively).

fMRI data acquisition

fMRI data were acquired in a 3 Tesla Siemens Prisma scanner with a 64-channel head coil (Siemens, Erlangen, Germany). Functional images were acquired using an accelerated T2*-

weighted gradient-echo echoplanar imaging (EPI) sequence (multiband factor = 2, repetition time (TR) = 2000 ms, echo time (TE) = 30 ms, flip angle (FA) = 75°, 62 slices with an isotropic voxel size of 2 × 2 × 2 mm) using GRAPPA (GRAPPA = 2). Each run consisted of 198 images (total duration = 6 minutes and 36 seconds). Moreover, a high-resolution T1-weighted structural scan with whole-brain coverage was performed for each participant (TR = 2000 ms, TE = 3.06 ms, inversion time (TI) = 1100 ms, FA = 9°, 192 slices and an isotropic voxel size of 1×1×1 mm). The structural scan was measured during the explanation of the tricks in the first set.

fMRI data preprocessing

Functional MRI images were preprocessed first by removing thermal noise from the magnitude EPI images using NORDIC, a PCA-based algorithm (Vizioli et al., 2021) in MATLAB. Then, we discarded the first five volumes of each run to allow for T1 equilibration effects. Using SPM12 we further performed motion correction (realigned to the first image), slice-time correction (using middle slices as reference) and co-registration to the structural scan. Finally, functional MRI data for whole-brain analyses were normalized to the Montreal Neurological Institute template brain (MNI152) and spatially smoothed with a Gaussian kernel of full width at half-maximum of 6 mm for univariate whole brain analyses. Region-of-interest (ROI) analyses were performed on unsmoothed data in native space. Moreover, for the generation of subject-specific ROIs, we generated inflated individual brain surfaces using Freesurfer 7.1.1 (Dale et al., 1999) using a dedicated docker container (https://hub.docker.com/r/freesurfer/freesurfer).

Univariate whole-brain data analysis

To investigate differences in neural activity during high-level VOE in the human brain, we performed whole-brain and ROI univariate analyses. We had three specific aims: 1) to investigate neural responses to violations of physical principles in naturalistic stimuli, comparable to previous reports using magic tricks (Danek et al., 2015; cf. Parris et al., 2009), 2) to investigate differential responses between different VOE (i.e., magic trick types) and 3) to investigate the role of prior knowledge in the perception of events violating physical principles by comparing responses to the very same videos before and after revelation.

For these analyses, we created two event-related general linear models (GLM) using the canonical hemodynamic response function and high-pass filtered data with a cut-off at 128s in SPM12. In the first GLM (aim 1), we modelled responses using only three regressors of interest in each run (magic, control, and unusual stimuli conditions). In a second, extended GLM, we investigated possible differential responses to different forms of violation of expectations (aim 2), using seven regressors of interest that modelled BOLD responses of the three magic types, three matching controls and unusual actions for each run. As in previous reports, we modelled individual trials within a regressor using discrete event times based on the moment of surprise (Danek et al., 2015; cf. Parris et al., 2009). The definition of the event times was done by averaging the

independent selection of a suitable frame done by two authors (VP and PRG). For magic videos, timing was decided based on when the "magic" (i.e., VOE) happened. For matching control videos, the corresponding time point was selected, i.e., the moment in which one would expect to see a magic event in the magic videos. For the videos showing unusual actions, timing was selected based on the onset of the unusual actions. Moreover, we included participants' response times, six movement parameters and a column of ones as nuisance regressors in both GLMs. We additionally computed the same analyses modelling the whole video durations (14 s) and report these in the supplementary Tables S11-13 and supplementary Figure S7). We used these models to address our three aims as follows.

Aim 1: to identify brain areas preferentially involved in signaling high-level VOE we compared responses to magic and matching control videos in all six pre-revelation runs (i.e., two runs from each object set) using the pooled regressors (from the first GLM). First-level contrasts $Magic_{pre} > Control_{pre}$ were used for second-level random-effect analyses with non-parametric permutations tests using the non-parametric mapping toolbox SnPM13 (Nichols and Holmes, 2002) for SPM12.

Aims 1 and 2: to investigate generic and violation-specific responses to high-level VOE before the explanation of the tricks, we used a second-level, 2 (magic, control) x 3 (appear, change, vanish) repeated measures ANOVA to conduct four conjunction analyses (Friston et al., 2005). To find common responses to all VOE, we contrasted each magic trick type with its corresponding control condition before the explanation of the tricks (i.e., $Magic_{A\_pre} > Control_{A\_pre}$) and used those contrasts to perform a conjunction (across all three VOE types) with a threshold of $p_{unc} \leq 0.001$ and a cluster threshold of k = 10. For each VOE type separately, we further performed a conjunction analysis on the responses of one VOE type against the others. For example, responses specific to an appearing object were investigated by means of the conjunction of the contrasts $Magic_{A\_pre} > Magic_{C\_pre} \cap Magic_{A\_pre} > Magic_{V\_pre}$.

Aim 3: to identify brain areas generally involved in the perception of magic and affected by prior knowledge we compared responses to magic after providing the explanation of the tricks $Magic_{post} > Control_{post}$ and additionally compared the interaction between the revelation and magic $(Magic_{pre} > Control_{pre}) > (Magic_{post} > Control_{post})$ using permutations tests. Finally, we compared the same contrasts for each magic type separately using the second, extended GLM, e.g., $(Magic_{A\_pre} > Control_{A\_pre}) > (Magic_{A\_post} > Control_{A\_post})$. Please note that these contrasts are controlling for potential time confounds.

All non-parametric permutation tests were performed using 5000 permutations, a cluster-forming threshold of $p_{unc} = 0.001$ and family-wise error correction (FWE, $\alpha = 0.05$). We report clusters that do not survive FWE-correction but exceed k = 10 voxels. Conjunction analyses based on parametric statistics were performed using an uncorrected threshold of $p \leq 0.001$ and a cluster threshold of k = 10. Brain areas in whole brain analyses were first identified using the atlasreader

toolbox (Notter et al., 2019), applying the Automated Anatomical Labeling atlas 3 (AAL3) (Rolls et al., 2020).

Univariate whole-brain control analyses

We performed two control analyses for the whole-brain univariate results. First, it is possible that the conjunction analyses examining differences between the magic trick types are confounded by differences in the visual content of the videos at specific time-points. This is because we are not only comparing videos showing “appearances”, “color changes” and “vanishes”, but also videos showing “red objects”, “blue objects” and “no objects” at a specific time point, respectively (see Figure 1C). To control for this possible stimulus-driven confound we compared responses between magic and control videos showing similar visual contents at specific time points. We contrasted responses to magic appearances with the control videos for vanishing tricks (as both videos show a “red object” at the specific moments), and magic vanishes with the control videos for tricks showing something appear (as both videos show no objects at the specific moments). No similar match was possible for the color-changing videos. Responses similar to that of the original contrasts, which used control videos showing similar actions, would be indicative that the observed differential activity reflects specific surprise responses and not differences in the visual content of the videos.

Second, we looked for condition-independent time effects in our whole-brain results. We performed a mixed-effects model using the average beta estimates of the magic tricks of each run, using the significant clusters of the prior-knowledge dependent contrast ($Magic_{pre} > Control_{pre}$) > ($Magic_{post} > Control_{post}$). We hypothesized that any prior-knowledge dependent modulation should show high activity in the runs before the explanation of the tricks and decreased activity afterwards. In contrast, any results driven by either a general or set-wise drop in alertness should show a gradual drop in the magic estimates across all runs of the experiment or within a set, respectively. Accordingly, we used four predictors in the mixed model to explain the 12 magic estimates based on their presentation order: one predictor modelled the revelation condition (i.e., 1, 1, -1, -1, etc.), one predictor modelled the run number (i.e., 5.5 to -5.5 in 12 steps) and two predictors modelled a decay in activity either before (i.e., 1, -1, 0, 0, etc.) or after the revelation of the tricks (i.e., 0, 0, 1, -1, etc.). We expect the last three regressors to account for any time-dependent variance.

ROI definition

For the ROI analyses, we defined 26 hypothesis-driven ROIs, separated into two groups.

First, we defined a set of 16 ROIs based on significant responses to magic videos from previous experiments (Danek et al., 2015; Parris et al., 2009). The authors from Danek et al., (2015) kindly provided us with the corresponding parametric maps from their study which were

used to guide the ROI definition. We combined individual labels from a multi-modal parcellation of the human cortex (Glasser et al., 2016) to define the following 14 frontal and parietal ROIs that showed increased activity when viewing magic tricks: a posterior part of the dorsal anterior cingulate cortex (pdACC), an anterior part of the dorsal ACC (adACC), ventral ACC (vACC), inferior frontal junction (IFJ), left inferior frontal sulcus (IFS), Brodman area 6 (BA6), inferior premotor subdivision (6r), 8BM, anterior insula (AI), anterior ventral insula (AVI), inferior temporal gyrus temporo-occipital division (PH), left BA 46, left BA 8 and left inferior parietal cortex (IPC). The remaining two subcortical ROIs (caudate nucleus and left amygdala) were defined using the Freesurfer automatic parcellation (Fischl et al., 2002). A detailed list of the parcellations used to define these ROIs is in the supplementary section *Surprise-related region of interest definition*.

Second, we used a probabilistic map of visual fields (Wang et al., 2015) to define ten visual ROIs: primary visual cortex (V1), secondary visual cortex (V2), V3, V3A, V3B, human V4 (hV4), lateral occipital and ventral complex (LO and VO, respectively), intraparietal sulcus (IPS) and frontal eye-fields (FEF). All ROIs were defined in native space.

We included these visual ROIs to use in the decoding of the VOE type (i.e., appear, change and vanishing, see below) and to test for differences evoked by the different VOE types and the effect of prior knowledge. For example, areas of the ventral visual cortex are known to be responsive to color (Bartels and Zeki, 2000), while the lateral occipital complex is responsive to objects (Grill-Spector et al., 2001). As predictive coding approaches predict feature-specific prediction errors in functionally specialized regions, we expect unexpected color changes to affect color-responsive ROIs (e.g., hV4, VO and PH), and unexpected object appearances to affect object-responsive areas LO and VO, in line with recent imaging evidence (Jiang et al., 2016; Richter et al., 2018; Stefanics et al., 2019). Early visual areas (V1, V2, V3) were included to investigate possible top-down effects of prior knowledge in lower-level areas, when comparing the exact same videos before and after revelations, while the parietal (IPS) and prefrontal ROIs (FEF) were included due to their involvement in top-down voluntary attention (Corbetta and Shulman, 2002). See Figure 5A for a depiction of all 26 ROIs in an exemplary subject.

Multivariate pattern analysis (MVPA)

Apart from the univariate analyses testing for net signal differences, we further wanted to investigate which areas of the brain carry pattern information about the different types of VOE (unexpected appearance, feature change, and omission). To do so, we performed a series of multivariate pattern analyses (MVPA) on the 26 hypothesis-driven ROIs and a control ROI.

For the decoding analyses we computed a GLM in which every trial (i.e., video presentation) was modeled as a separate regressor to increase the number of data points for training and testing. All analyses were performed using a shrinkage linear discriminant analysis (LDA) on the de-meaned beta estimates of the individual trials (by the mean over all estimates, i.e., all trials, within each voxel) using the Python (version 3.8.13) package scikit-learn's class

LinearDiscriminantAnalysis (Pedregosa, 2011). To examine if any of the ROIs contained information about the different types of VOE (appear, color change, and vanish) we trained and tested a shrinkage LDA to predict the VOE types following a three-fold cross-validation scheme to ensure generalization across objects. We trained on the data of two objects (i.e., estimates from two sets, 48 trials) and tested on the third object (i.e., estimates from the third set, 24 trials). Significance testing of the decoding accuracies was done using a permutation analysis (1000 permutations) implementing the max statistic correction to correct for multiple comparisons (Nichols and Holmes, 2002). A control ROI (third ventricle) was included in the analysis, which should carry no information and thus reflect chance level.

Decoding analyses were performed separately for data before and after explanations of the magic tricks. Permutation-based corrected significance thresholds were 36.98% and 36.92% before revelation and after revelation, respectively. As both analyses were conducted using estimates based on the very same videos, we hypothesized that any significant difference in decoding accuracies between data before and after revelation would be indicative of decodable prior-knowledge dependent surprise signals. We tested for differences in decoding accuracies using paired t-tests between decoding using pre-revelation data and decoding using post-revelation data, only in those ROIs that showed significant decoding (corrected) using pre-revelation data.

Additionally, as an exploratory approach we performed a similar whole-brain searchlight analysis using a sphere to decode the magic types (4 mm radius), separately for data before and after explanation of the tricks using the SearchLight class implemented in nilearn (Abraham et al., 2014). The searchlight whole-brain accuracy maps were spatially smoothed with a 4 mm Gaussian kernel. A permutation-bootstrap hybrid method (in which each randomly generated accuracy map was also smoothed with a 4 mm Gaussian kernel) was used for significance testing and correction for multiple comparisons (Stelzer et al., 2013) using custom-made Python code.

Behavioral data

To test for differences in surprise ratings between videos before and after the explanation of the tricks we performed a 2 (before/after explanation) x 3 (magic, unusual and control videos) repeated measures ANOVA. We expected to see higher surprise ratings for magic videos compared to control videos and higher ratings for magic videos before compared to after the revelation of the methods. We also wanted to test if videos of magic and of unusual actions led to similarly high surprise ratings. Only this would allow us to use the unusual action videos as secondary comparison points for magic. Moreover, we tested for differences in surprise ratings of the magic videos for different objects and VOE types in 2 x 3 repeated-measures ANOVAs (rmANOVA) with the factors revelation (before/after) and object (ball/card/pencil) or VOE type (appear/change/vanish), respectively.

Eye tracking

Data acquisition and preprocessing. Gaze positions were measured using an MR-compatible Eyelink 1000 (SR-Research, Ottawa, Canada) positioned at the rear end of the scanner bore at a 1000 Hz recording rate. Calibration of the eye tracker was performed at the beginning of the experiment and drift correction was performed at the beginning of each run. If necessary, re-calibration was performed at the beginning of a new run.

Eye tracking data (i.e., gaze position, blinks, and saccades) from 23 participants were analyzed. Data from one participant was excluded because of technical problems during data acquisition. Monocular gaze path data (x, y coordinates) were cleaned by removing values 150 ms before and after blinks, linearly interpolating the missing data (with an interpolation limit of 500 ms) and downsampling the data from 1000 Hz to 25 Hz (i.e., the framerate of our videos). Identification of blinks and saccades was performed automatically using the Eyelink online parser with default parameters (saccade detection threshold was 22 degrees/s).

Eye tracking data was used to examine if fMRI responses could be confounded by systematic differences in gaze traces, saccades, or blinks during viewing of the videos. We focused our eye tracking analyses to values within a −1 to +2 s window centered around the event time from each video. We performed two types of tests. First, we used rmANOVAs to test for differences in blink or saccade numbers across conditions. Second, we used correlation analyses to test for differences in gaze traces.

Gaze trace analysis. To test for differences in gaze traces before and after the revelation of the methods used in the magic tricks, we correlated the x and y positions of each video presentation (eight presentations per video – two per run, two runs pre and post revelation) within a subject, for each video separately (resulting in two 8 × 8 correlation matrices per video – one for x, one for y.). Then, the correlation coefficients were transformed using the Fisher-z transformation and averaged for the x and y traces. We pooled all values from comparisons *between* presentations before and after the revelation of the tricks (i.e., the bottom left quadrant of the matrix) and those *within* presentation before and after the revelation (see Figure 2E for an example) of the tricks within a subject. The pooled correlation coefficients were then compared using paired t-tests.

Blink and Saccade analysis. Finally, to test for differences in blinks and saccades as a factor of video condition, prior knowledge, and VOE, we compared the mean number of blinks and saccades using two rmANOVAs, for blinks and saccades separately. The first rmANOVA had video (magic and control) and revelation (before and after) condition as factors. The second rmANOVA used magic data and had the VOE types (appear, color change, and vanish) and revelation (before and after) as factors.

Inference statistics

Effect sizes for repeated measures ANOVAs and paired tests are presented as partial eta squared (η^2) and Cohen's *d*, respectively. Sphericity of rmANOVAs was tested using the Mauchly test. If sphericity was violated, degrees of freedom were adjusted using the Greenhouse-Geisser correction, the corresponding ε-correction factor is provided. Normality of data was assessed using the Shapiro-Wilk test (for paired tests and post-hoc tests). In case that data was normally distributed, we performed paired t-tests, otherwise we performed non-parametric Wilcoxon signed-rank tests instead. Correction for multiple comparison was performed using a step-down Holm-Bonferroni correction. Please note that in the post-hoc tests we corrected for each hypothesis separately (i.e., p-values for post-hoc tests of one factor are corrected independent of another factor or an interaction). In general, corrected p-values (p_{corr}) are reported in text, uncorrected p-values (p_{unc}) are reported in the corresponding tables. The threshold for statistical significance was set to 0.05 for all tests.

Results

Behavioral surprise ratings

Behavioral data was first tested for differences in surprise ratings for video condition (magic, control, and unusual videos) and revelation condition (before and after revelation) using a 2 x 3 rmANOVA (see Figure 2A). In sum, this analysis revealed that magic tricks were perceived as more surprising than the control videos, and that surprise ratings dropped after explanation of the tricks. In detail: the analysis revealed significant main effects for both factors (video condition: $F(2,46) = 57.494$, $p_{unc} < 0.001$, $\eta^2 = 0.478$, $\varepsilon = 0.932$, revelation: $F(1,23) = 103.242$, $p_{unc} < 0.001$, $\eta^2 = 0.211$, $\varepsilon = 1$) and interaction: $F(2,46) = 43.081$, $p_{unc} < 0.001$, $\eta^2 = 0.134$, $\varepsilon = 0.641$). As expected, post-hoc Wilcoxon signed-rank tests revealed that magic videos were more surprising than control and unusual videos, before ($W = 1$, $p_{corr} < 0.001$, Cohen's $d = 4.07$ and $W = 8$, $p_{corr} < 0.001$, Cohen's $d = 2.06$, respectively) and after explanation of the tricks ($W = 5$, $p_{corr} < 0.001$, Cohen's $d = 1.55$ and $W = 26$, $p_{corr} < 0.001$, Cohen's $d = 0.812$, respectively), and that all video conditions were more surprising before compared to after the revelation (all three video conditions: $p_{corr} <= 0.001$). Unusual videos were more surprising than control videos pooled across runs ($W = 42$, $p_{corr} = 0.006$, Cohen's $d = 0.59$), but significantly so only in the first two runs ($W = 26$, $p_{corr} = 0.002$, Cohen's $d = 0.71$) (see Supplementary Table S1 for a detailed report of all post-hoc tests). Because of the large and significant differences in surprise ratings between the magic and unusual videos that preclude a meaningful comparison of corresponding neural responses, we report here no further analysis of the data from the unusual videos. For completeness, we show corresponding contrasts in the supplementary results (see supplementary Figure S3 and Tables S3-6).

Average surprise ratings of magic videos were consistently high before the explanation of the tricks (all group means > 3) and decreased afterwards (all group means < 2.5) (see Figure 2A, B and C). We further performed two rmANOVAs using only ratings from magic videos, to test for possible differences in VOE types and the objects used. Both rmANOVAs included the revelation condition as a factor and showed a significant decrease in surprise rating after the revelation of the tricks (as expected from the previous analysis). We additionally found a significant main effect for the magic type ($F(2,46) = 13.8$, $p_{unc} < 0.001$, $\eta^2 = 0.032$, $\varepsilon = 0.848$) and an interaction of magic type and revelation condition ($F(2,46) = 11.02$, $p_{unc} < 0.001$, $\eta^2 = 0.018$, $\varepsilon = 0.795$) (see Figure 2B). Post-hoc tests showed that appearances were rated less surprising than color changes and disappearances pre revelation ($W = 12$, $p_{corr} < 0.001$, Cohen's $d = -0.835$ and $W = 35.5$, $p_{corr} = 0.002$, Cohen's $d = -0.631$, respectively). No difference between color change and vanish magic tricks was observed (see Supplementary Table S2). No main effect for objects nor an interaction of object and revelation condition were found (all *F-values* < 2, all $p_{unc} > 0.15$ and $\eta^2 < 0.012$) (see Figure 2C).

Eye-tracking results

Eye-tracking data were tested for systematic differences in gaze traces, saccades, and blinks during viewing of the videos (see also Supplementary section *Eye-tracking results* and Figure S2). First, for each magic video we computed the correlation of gaze path between video presentations in all subjects. We then averaged the transformed values (using Fisher-z transformation) of correlations between pre- and post-revelation presentations and within pre- and post-revelation presentations and tested for differences using paired t-tests. Gaze traces between pre- and post-revelation runs were similar (see an example visualization in Figure 2D) and no significant difference of correlations of gaze traces was observed (all p_{unc} > 0.2) (see Figure 2E for an example). Moreover, the number of saccades and blinks around the VOE times were similar between experimental conditions and only revealed small differences we deem unlikely to have affected the imaging results.

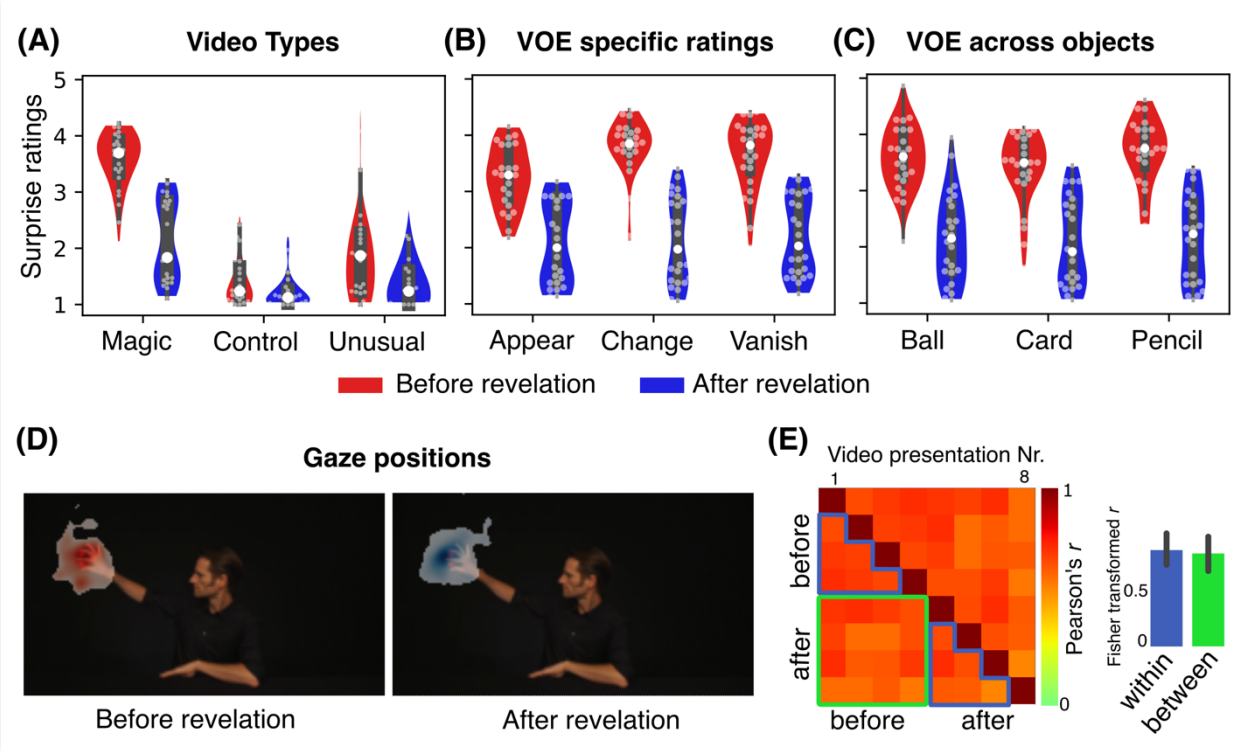


Figure 2. Shown are behavioral surprise ratings separated for the different video types (**A**), for magic videos across VOE types (**B**) and across objects (**C**). Surprise ratings before the revelations of the magic tricks (red) were consistently higher than after the revelations (blue). **D**, Exemplary group gaze positions for the same magic video in the moment a ball appears, before (left) and after (right) the revelation of the trick. **E**, Correlation matrix for gaze positions around the moment of magic (-1 s and +2 s) from an exemplary video. To test for differences between gaze paths before and after the revelation of the tricks, we compared the average Fisher z-transformed correlations between gaze paths of all

combinations of presentations (from presentation Nr. 1 to Nr. 8) *between* pre- and post-presentations (marked green) and those *within* pre- and *within* post-presentations (marked blue), as shown in the bar plot.

Univariate whole-brain analysis

Whole-brain surprise responses

To investigate neural correlates of high-level VOE when viewing seemingly impossible events, we first compared whole-brain responses to the magic and matched control videos before the revelation of the magic tricks ($Magic_{pre}$ > $Control_{pre}$). This contrast revealed several clusters of activity in frontal and parietal cortices, as well as subcortical areas, such as the caudate nucleus, largely in line with previous studies (Danek et al., 2015; Parris et al., 2009) (see Figure 3A and Table 1). In particular, large clusters of activity were observed in the medial part of Brodmann area 8 (preSMA), the dorsal and ventral anterior cingulate cortex (dACC and vACC), caudate nucleus (CN), and the posterior parietal cortex (PPC, especially the superior parietal lobe and the precuneus). No lower-level sensory area was differentially modulated in view of the unexpected events.

Generic surprise responses

To specifically test for generic surprise responses in the brain showing a significant involvement of all three types of VOE, we performed a conjunction analysis combining all different VOE types before the revelation of the tricks ($Magic_{A\_pre}$ > $Control_{A\_pre}$ ∩ $Magic_{C\_pre}$ > $Control_{C\_pre}$ ∩ $Magic_{V\_pre}$ > $Control_{V\_pre}$). While no cluster survived FWE correction, using a voxel-wise uncorrected threshold of p ≤ 0.001 and k = 10 we found small clusters in the dorsal anterior cingulate cortex (dACC) bilaterally and posterior parietal cortex (precuneus) revealing generic responses (see Figure 3B, left and Table 1). Also, using a more liberal threshold of p_{unc} ≤ 0.005 (and k = 10), we further observed generic activity in the vACC/mPFC.

Table 1. Significant clusters of activity from the whole-brain contrast comparing responses between magic videos and matched controls ($Magic_{pre}$ > $Control_{pre}$, thresholded at p ≤ 0.001 and k = 10, uncorrected) and the conjunction analysis testing for common areas involved in the processing of violation of expectations (thresholded at p_{unc} ≤ 0.001 and k = 10, uncorrected). P-values show cluster statistics. k = cluster size, T = t statistic at peak voxel, x, y, z = peak voxel MNI coordinates [mm]. dAAC: dorsal anterior cingulate cortex; PFC: prefrontal cortex; vACC: ventral ACC; mPFC: medial PFC; PCC: posterior cingulate cortex; DLPFC: dorso-lateral prefrontal cortex; preSMA: supplementary motor area.

Brain region	AAL atlas labels	p(FWE)	p(unc)	k	T	x	y	z
$Magic_{pre}$ > $Control_{pre}$								
Posterior parietal cortex	**Occipital_Mid_L**	**0.0004**	<0.0001	1871.0	7.16	-34	-82	34
	Precuneus_R	-	-	-	6.87	4	-68	52
	Precuneus_L	-	-	-	6.48	-10	-68	44
dAAC	**Cingulate_Ant_L**	**0.0048**	0.0005	886.0	6.46	-6	30	20

	Cingulate_Ant_R	-	-	-	6.05	4	32	24
	Cingulate_Mid_R	-	-	-	5.51	2	24	32
L. Parieto-occipital sulcus	**Cuneus_L**	0.0594	0.0082	162.0	6.18	-14	-62	22
	Calcarine_L	-	-	-	4.21	-12	-60	14
R. Intraparietal sulcus	**Occipital_Mid_R**	0.1752	0.0284	74.0	5.53	36	-80	32
L. anterior PFC	**Frontal_Sup_2_L**	0.1982	0.0332	67.0	5.41	-26	58	4
L. anterior intraparietal	**Postcentral_L**	0.0548	0.0074	172.0	5.2	-46	-36	54
	Parietal_Inf_L	-	-	-	4.47	-42	-38	44
vACC/mPFC	**Cingulate_Ant_L**	0.0548	0.0074	172.0	5.13	-4	48	2
L. Superior frontal gyrus	**Frontal_Sup_2_L**	0.0522	0.007	179.0	5.02	-22	12	58
	Frontal_Sup_2_L	-	-	-	3.7	-22	0	52
L. anterior middle frontal gyrus	**Frontal_Mid_2_L**	0.1722	0.028	75.0	5.01	-34	54	14
L. PCC	**no_label**	0.5014	0.1159	25.0	5.01	10	-36	2
L. Thalamus	**Thalamus_L**	0.5456	0.1334	22.0	4.86	-6	-6	10
L. Hippocampus	**no_label**	0.5158	0.1214	24.0	4.76	-18	-22	-10
L. Caudate nucleus	**Caudate_L**	0.5946	0.1558	19.0	4.73	-12	16	2
L. Postcentral gyrus	**Postcentral_R**	0.1134	0.0173	103.0	4.71	52	-16	54
	Postcentral_R	-	-	-	4.07	52	-18	46
	Postcentral_R	-	-	-	3.72	46	-22	42
R. Superior frontal gyrus	**Frontal_Sup_2_R**	0.2394	0.0417	57.0	4.61	28	10	62
L. PCC	**no_label**	0.2832	0.0511	49.0	4.48	-2	-30	26
R. Thalamus	**Thalamus_R**	0.7188	0.2273	13.0	4.17	12	-8	12
L. Fusiform gyrus	**Fusiform_L**	0.6338	0.1754	17.0	4.16	-24	-54	-16
	Cerebelum_6_L	-	-	-	3.75	-24	-62	-18
L. Ventral occipital cortex	**Lingual_L**	0.7188	0.2273	13.0	4.1	-2	-78	-8
L. Anterior intraparietal	**Parietal_Inf_L**	0.5306	0.1277	23.0	3.84	-40	-50	50
L. DLPFC	**Frontal_Mid_2_L**	0.762	0.2623	11.0	3.81	-36	26	32
Conjunction of *Magic*$_{pre}$ − *Control*$_{pre}$ per VOE type								
dACC/preSMA	**Supp_Motor_Area _L**	0.722	0.096	38	4.02	-4	16	50
Precuneus	**Precuneus_L**	0.885	0.162	26	3.99	-8	-68	48
	Precuneus_R	0.722	0.096	38	3.94	8	-68	48
dACC	**Cingulum_Mid_L**	0.966	0.254	17	3.7	-6	18	40
	Cingulum_Mid_L	-	-	-	3.31	-4	26	38
R. Postcentral gyrus	**Precentral_R**	0.986	0.318	13	3.5	46	-14	54

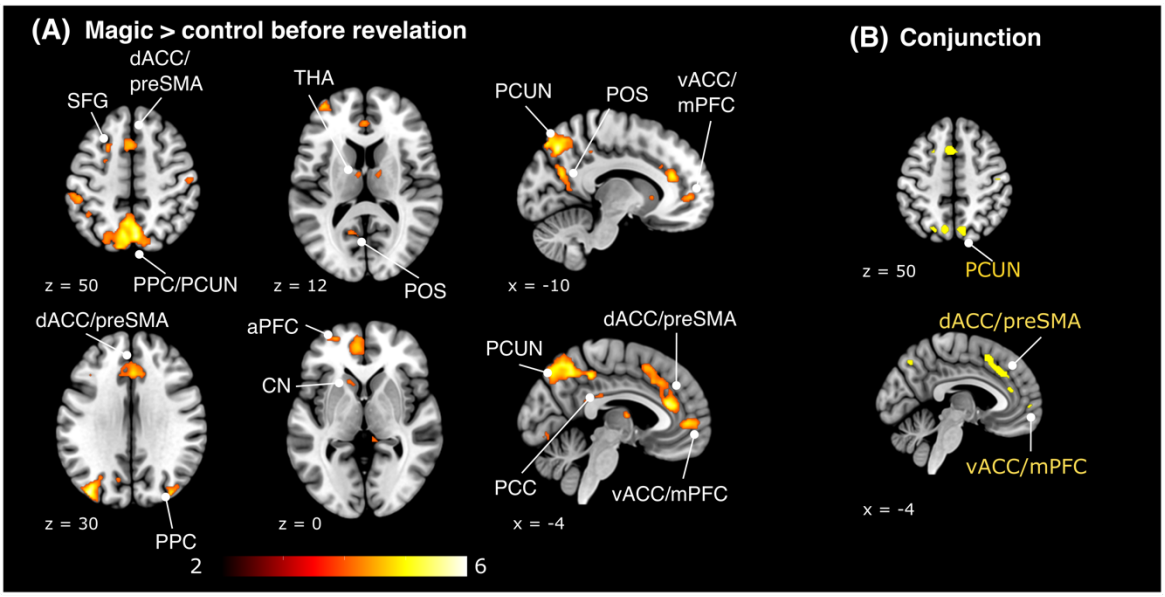
Specific surprise responses

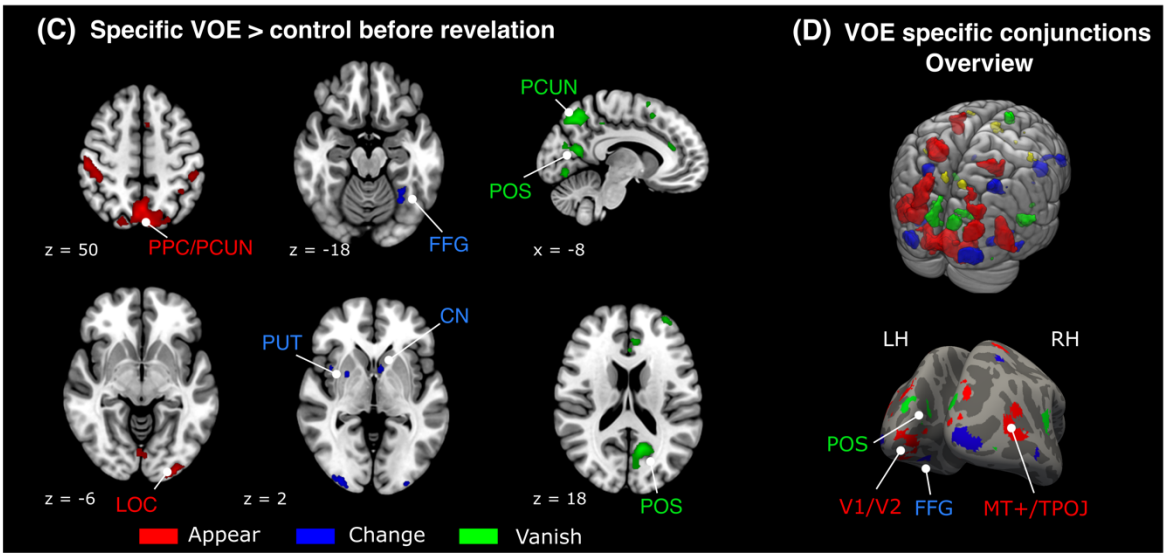
After establishing what areas are generally involved in the processing of violation of expectations (i.e., commonly active in seemingly impossible appearances, disappearances, and color changes), we looked for VOE type specific differential activity in the brain (e.g., areas responsive to something unexpected appearing but not disappearing or changing color). Beyond the systematic generic activation of frontoparietal areas described above, all three types of VOE

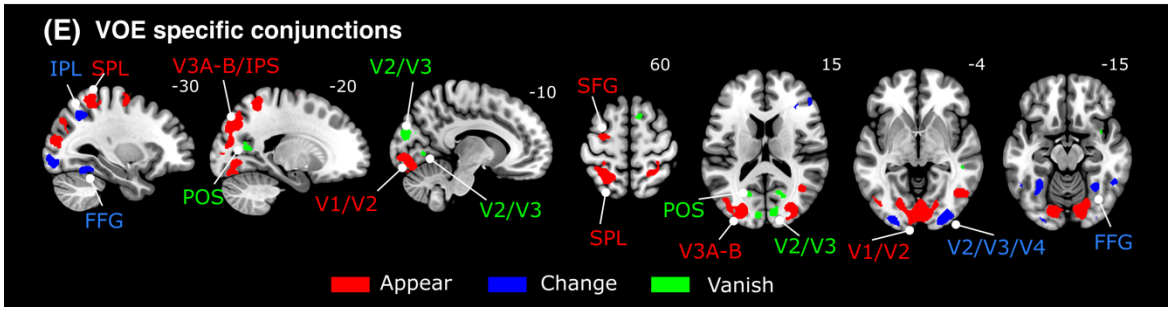
evoked differential responses in posterior visual areas, but in a segregated and even partially opposite manner (see Figure 3C and Supplementary Table S8). For example, while appearances and color changes evoked an increase of activity in ventral visual areas, the disappearance of objects reduced activity in overlapping regions. To better visualize this diverse modulation of sensory areas by the different VOE, we tested which areas were significantly more activated by one type of VOE than by the other two using a conjunction test (e.g., test for appear-specific responses: $Magic_{A\_pre} > Magic_{C\_pre} \cap Magic_{A\_pre} > Magic_{V\_pre}$) (see Figure 3D and E and Supplementary Table S7). As hypothesized, activity related to objects appearing were observed in early visual areas and in higher-level visual areas of the lateral occipito-temporal cortex (peak coordinates: x = 30,y = -88, z = 22 and x = -22, y = -78, z = 44), whilst activity related to color changes were observed specifically in color-responsive ventral areas of the fusiform gyrus (FFG, peak coordinates: x = -30, y = -50, z = -16 and x = 32, y = -54, z = -14). Finally, vanishing objects evoked significant responses – among others – in the anterior parts of the calcarine sulcus, close to the parieto-occipital sulcus (peak coordinates: x = -20, y = -64, z = 12 and x = 18, y = -76, z = 8).

Control analysis for visual content

As these VOE type-specific patterns of activity are located predominantly in visual processing areas, they are potentially related to general visual differences between the conditions tested. The compared videos are not only showing “appearances”, “color changes” and “vanishes”, but also “red objects”, “blue objects” or “no objects”, respectively. To test if these VOE type-specific responses are confounded by different visual input, we performed a control analysis comparing neural responses between 1) videos showing red objects, either as the product of a magic trick (appearances) or as a control to other tricks (vanishes) $(Magic_{A\_pre} > Control_{V\_pre})$ and 2) between videos showing no object, either as the product of a magic trick (vanishes) or as a control to the other tricks (appearances) ($Magic_{V\_pre} > Control_{A\_pre}$). Differential responses in posterior visual areas to these control contrasts were similar to the results of the corresponding conjunction analyses (for appearances and vanishes), suggesting that violation-specific responses observed in posterior visual areas are unlikely to be driven merely by visual content (see overlays in Supplementary Figure S5). However, if these signals in visual areas reflect specific prediction errors based on different VOE, we would additionally expect them to be modulated by prior knowledge and to show decreased responses after the explanation of the tricks.


(A) Magic > control before revelation
dACC/
preSMA
SFG
THA
PCUN
POS
vACC/
mPFC
z = 50
PPC/PCUN
z = 12
POS
x = -10
dACC/preSMA
aPFC
CN
PCUN
dACC/preSMA
PCC
vACC/mPFC
z = 30
PPC
z = 0
x = -4
2
6
(B) Conjunction
z = 50
PCUN
dACC/preSMA
vACC/mPFC
x = -4


(C) Specific VOE > control before revelation
PCUN
POS
z = 50
PPC/PCUN
z = -18
FFG
x = -8
CN
PUT
z = -6
LOC
z = 2
z = 18
POS
Appear
Change
Vanish
(D) VOE specific conjunctions
Overview
LH
RH
POS
V1/V2
FFG
MT+/TPOJ


(E) VOE specific conjunctions
IPL
SPL
V3A-B/IPS
V2/V3
-30
-20
-10
SFG
60
15
-4
-15
FFG
POS
V1/V2
V2/V3
SPL
POS
V3A-B
V2/V3
V1/V2
V2/V3/V4
FFG
Appear
Change
Vanish

1

Figure 3. A, shown are active regions during viewing magic tricks compared to matched control videos before the participants knew how the tricks were performed (thresholded at $p_{unc} < 0.001$ and k =10, uncorrected). **B**, shown are results from the conjunction analyses testing for generic activity (left, yellow) (at $p_{unc} < 0.005$). Only the dorsal ACC and preSMA and the precuneus (PCUN) revealed generic responses to high-level VOE at $p_{unc} < 0.001$. **C**, VOE specific activity revealing several posterior visual areas responsive to appearances (red), color changes (blue) and objects disappearing (green). **D**, overview of all conjunctions results in the MNI152 template volume (upper) and of the VOE-specific responses projected on an average surface (fsaverage) (lower). **E**, shown are VOE-specific conjunction results. **Abbreviations:** ACC: anterior cingulate cortex; dACC: dorsal ACC; vACC: ventral ACC; CN: caudate nucleus; aPFC: anterior prefrontal cortex; FFG: fusiform gyrus; IPL: interior parietal lobe; IPS: intraparietal sulcus; MT+: motion area MT; mPFC: ventral prefrontal cortex; PCC: posterior cingulate cortex; PCUN: precuneus; POS: parieto-occipital sulcus; PPC: posterior parietal cortex; SFG: superior frontal gyrus; SMA: supplementary motor area; SPL: superior parietal lobe; THA: thalamus; TPOJ: temporo-parietal-occipital junction. LH: left hemisphere; RH: right hemisphere.

Prior-knowledge dependent whole-brain responses

To investigate the effect of prior knowledge on brain responses, we provided participants with the methods behind the magic tricks. Surprisingly, neural activity after explanation of the tricks ($Magic_{post} > Control_{post}$) were similar to that before the explanation of the tricks (see Figure 4A and Supplementary Table S10). Thus, areas related to the processing of surprising events remained significantly active in view of VOE also after the explanation of the tricks (i.e., dACC, caudate nucleus, anterior insula). The interaction contrast comparing surprise responses before and after providing the explanations ($Magic_{pre} > Control_{pre}$) > ($Magic_{post} > Control_{post}$) revealed only a small number of areas showing prior-knowledge dependent modulations (see Figure 4B and C and Table 2). We observed higher activation in a large cluster of the medial prefrontal cortex (mPFC) and ventral ACC (peak coordinates: x = 2, y = 46, z = -10) and right posterior cingulate cortex (PCC) with FWE-correction (peak coordinates: x = 6, y = -46, z = 8). These regions hence decreased magic-related activity after the revelation. These areas overlap with and constitute of subset of the activity observed with $Magic_{pre} > Control_{pre}$ before the revelation of the tricks.

Interestingly, the observed decrease of activity in the vACC/mPFC and PCC after revelation of the tricks coincides with the midline core areas of the default mode network (DMN), whilst parietal areas of the dorsal attention network (DAN) showed increased activity. This indicates that our findings are unlikely a result of a decrease in attention, as the DAN is known to direct top-down attention (Corbetta and Shulman, 2002). A visualization of these patterns of activity, together with the DMN and DAN is shown in Figure 4C.

No prior-knowledge dependency of trick-specific modulations

While prior-knowledge driven signals overlapped with neural activations to unspecific VOE ($Magic_{pre} > Control_{pre}$), none of the previously observed trick-specific visual areas showed a modulation as factor of prior knowledge (even when using a more liberal threshold of $p < 0.001$ uncorrected, and a cluster size of k = 10, see Table 2 and Figure 4B and C). Further, the contrasts investigating the effect of prior knowledge for each VOE type separately revealed similar response

patterns to those reported in the main contrasts (see supplementary Figure S4). This indicates that visual areas responsive to a specific VOE (e.g., an object changing color) were not modulated by prior knowledge.

Controlling for time effects in net responses

To rule out possible time confounds in results comparing responses before and after revelation of the tricks, we inspected how the individual values changed across time as a control analysis. We extracted betas estimates corresponding to all magic and matched control presentations from the suprathreshold clusters from the prior-knowledge interaction contrast, averaged them over subjects and calculated the same contrast within each run separately. We then ran a mixed-effects model on the contrast values with four predictors: a pre-post predictor (i.e. 1, 1, -1, -1 for each set), a constant decay predictor over all 12 fMRI runs (i.e. 5.5 to -5.5 in 12 steps) and two time-decay predictors for values before and after revelation separately (i.e. 1, -1, 0, 0 and 0, 0, 1, -1, for each set). We hypothesized that areas showing prior-knowledge dependent activity should have a significant pre-post predictor. The pre-post predictor was significant in all clusters (all p_{corr} < 0.05). However, in some clusters a significant amount of variance was also explained by the time-decay regressors, showing that the activity in some clusters had a contribution of time. Yet, the fact that in all clusters the pre-post regressor remained significant despite inclusion of the time-regressor shows that knowledge-dependent effects were true (see supplementary Figure S6 and Table S9).

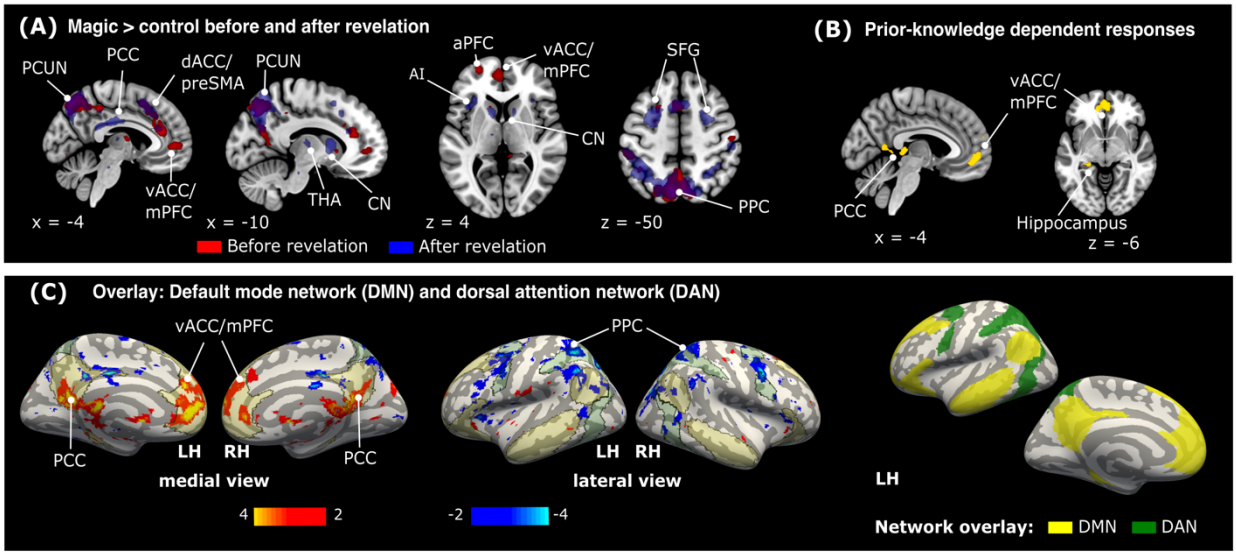


Figure 4. Prior knowledge dependent modulation of brain responses. **A,** overlay of *Magic* > *Control* before (red) and after (blue) explanation of the magic tricks. Both contrasts reveal a consistent activation of surprise related areas, such as the dorsal ACC (dACC), caudate nucleus (CN), and posterior parietal cortex (PPC). **B,** the difference of both contrasts (i.e., the interaction between video condition and revelation) revealed only few areas that were significantly more active before the explanation of the tricks: the posterior cingulate cortex (PCC), the ventral ACC/medial prefrontal cortex (vACC/mPFC) and left Hippocampus. Threshold at p_{unc} < 0.001 and k = 10. **C**, shown are the interaction testing for prior-knowledge dependent responses (thresholded from t = 2 to t = 4, red to yellow and t = -2 to t = -4, dark blue to light blue) and the default mode network (DMN) and the dorsal attention network (DAN) as transparent overlay projected on to an average brain surface (fsaverage). The network overlays are provided by (Yeo et al., 2011). **Abbreviations:** AI: anterior insula; aPFC: anterior prefrontal cortex; LH: left hemisphere; PCUN: precuneus; preSMA: supplementary motor area; RH: right hemisphere; SFG: superior frontal gyrus; THA: thalamus.

Table 2: Results of contrasts comparing neural responses to VOEs before and after explanation of the tricks (thresholded at p_{unc} < 0.001 and k = 10, uncorrected). P-values show permutation-based cluster statistics. K = cluster size, T = t statistic at peak voxel, x, y, z = peak voxel MNI coordinates [mm]. ACC: anterior cingulate cortex; vACC: ventral ACC; mPFC: medial prefrontal cortex; POS: parieto-occipital sulcus; PCC: posterior cingulate cortex; dACC: dorsal ACC.

Brain region	AAL atlas labels	p(FWE)	p(unc)	k	T	x	y	z
vACC/mPFC	**Frontal_Med_Orb_R**	**0.0068**	0.0007	538.0	5.72	2	46	-10
	Frontal_Med_Orb_L	-	-	-	4.96	-10	42	-6
R. Cuneus (V3)	**Cuneus_R**	0.25	0.0385	52.0	5.23	8	-86	34
L. POS	**Precuneus_L**	0.1082	0.014	101.0	4.99	-8	-56	8
L. PCC	**Calcarine_L**	-	-	-	4.31	-6	-48	4
R. PCC	**Precuneus_R**	**0.0424**	0.005	185.0	4.91	6	-46	8
	no_label	-	-	-	4.59	-4	-30	6
	no_label	-	-	-	4.56	8	-36	4
L. Temporal pole	**Temporal_Pole_Sup_L**	0.7558	0.2264	12.0	4.67	-32	14	-28
L. dACC	**Cingulate_Ant_L**	0.6462	0.1596	17.0	4.53	-10	38	14
L. Hippocampus	**Hippocampus_L**	0.665	0.1695	16.0	4.21	-22	-32	-6
L. Hippocampus	**Hippocampus_L**	0.5714	0.1262	21.0	4.14	-24	-18	-14

Univariate ROI analysis

To complement the whole-brain analysis, we defined a set of 26 hypothesis-driven regions-of-interest (ROIs) from which we extracted and analyzed parameter estimates (10 visual and 16 surprise-related ROIs based on prior literature) from unsmoothed data in native space and tested them with paired tests (correcting for the number of ROIs). Overall, our ROI-based analyses confirmed our above whole-brain findings. None of the lower-level visual cortices showed a significant increase of neural responses during the magic compared to the matched control condition before the explanation of the tricks (all p_{unc} > 0.24). In contrast, significant responses to magic were observed in the visual parietal ROI IPS (t(23) = 2.746, p_{unc}= 0.012, Cohen's d = 0.53),

as well as in several higher-level surprise-related ROIs, especially in adACC ($t(23) = 5.32$, $p_{corr} = 0.001$, *Cohen's d* = 0.917), as well as in vACC, pvACC, BA6, BA46 and caudate nucleus (all $p_{unc} < 0.05$, see detailed results in supplementary Table S14-17). Only two ACC ROIs (adACC and vACC) and 8BM (directly superior to the adACC) showed a decrease of activity after the revelation of the tricks ($t(23) = 3.43$, $p_{unc} = 0.002$, Cohen's $d = 0.41$, $W = 79$, $p_{unc} = 0.042$, Cohen's $d = 0.42$ and $t(23) = 2.54$, $p_{unc} = 0.01$, Cohen's $d = 0.41$, respectively). None of the visual ROIs showed a similar response pattern. In contrast, VOE-specific responses were largely constrained to visual ROIs (see supplementary section *univariate ROI results*).

Multivariate pattern analysis

Our univariate analyses revealed generic responses (modulated by prior knowledge) in fronto-parietal areas, while showing trick-specific responses in posterior sensory areas (unaffected by prior knowledge). However, differences in specific VOE and prior-knowledge modulations might also be reflected in activation patterns and not only in net signal differences. Accordingly, we complemented our univariate analysis by a multivariate pattern analysis to investigate if specific information about VOE types were present in the activity patterns of our surprise-related fronto-parietal ROIs (which showed only generic responses to magic). Further, we performed this analysis separately for data before and after the explanation of the tricks, to investigate if posterior sensory areas (which showed trick-specific effects) might show a difference in decoding accuracies before and after revelation of the tricks.

We followed a three-fold cross-decoding approach to generalize each magic type across objects: we trained a linear classifier to classify specific VOE types, appear (A), color change (C), and vanishes (V), based on the estimates of the magic videos from two objects (48 trials, from two sets) and tested on the estimates of the remaining object (24 trials, from one set).

Visual ROIs

Decoding of the specific VOE types across objects was possible in all posterior visual ROIs (V1, V2, V3, hV4, V3A, V3B, LO, VO, IPS) before the explanation of the magic tricks (all corrected p-values < 0.001; corrected using a permutation maximal statistic, Nichols and Holmes, 2002). After revelation of the method behind the magic tricks, decoding accuracies significantly dropped in most ROIs (V1, V2, V3, LO and IPS) (all $p_{corr} <= 0.05$; Holm-Bonferroni corrected for the number of visual ROIs), being below threshold in LO and IPS (see Figure 5B and supplementary Table S20). Additionally, we found uncorrected significant differences in hV4 and V3B ($p_{unc} = 0.05$ and $p_{unc} = 0.016$, respectively).

Surprise related ROIs

In contrast, no surprise-related ROI (nor the control ROI) showed significant above-chance decoding accuracies between VOE types using the permutation-max statistic for correction, except for the PH ROI (p_{corr} < 0.001), an area located in the inferior temporal sulcus (temporo-occipital division), using data before revelation of the tricks. Uncorrected significant above chance decoding (chance level = 33%) was observed in IFJ, AI, pdACC, BA6, BA8, and BA46 using data before revelation. Differences in decoding accuracies before and after revelation were observed only in PH (p_{corr} = 0.005, Holm-Bonferroni corrected for the number of ROIs that could significantly decode before revelation) (see Figure 5B and supplementary Table S20) and in BA8 (corrected for the number of ROIs that could *not* significantly decode the VOE type using pre-revelation data, i.e., 16 ROIs) (see Figure 5B and supplementary Table S21).

Therefore, while frontoparietal areas showed a generic and surprise-dependent involvement in processing VOE in the univariate analyses, they carried no or only weak information (i.e., in BA8) as to what exactly happened. In contrast, information about specific VOE were observed in all posterior sensory areas across the visual hierarchy. An in-depth view of the informative ROIs is shown in the confusion matrices in Figure 5C. They reveal that appearances and color-changes were consistently decodable above chance and that most confusions happened between them and less so with vanishing events.

Searchlight analysis

To complement the ROI decoding analysis that used hypothesis-driven visual and surprise-related ROIs, we performed a whole-brain searchlight analysis to better visualize the results and explore if additional brain areas differentially respond to specific VOE. As in the ROI decoding analysis, we trained linear classifiers to classify specific VOE (i.e., appear, change, vanish) in a three-fold approach separately for both, magic estimates before and after revelation of the tricks. Results of the whole-brain searchlight analysis revealed that only posterior visual areas of the brain could significantly decode the magic type in both conditions (using a permutation-bootstrap hybrid correction method, Stelzer et al., 2013). As shown in Figure 5D, significant decoding was possible in large areas of the visual cortex, including most of the occipital cortex and small parts of the temporal and parietal cortex. Crucially, decoding accuracies before the explanation of the magic tricks were significant in more voxels and larger clusters compared to after explanation. Significant decoding before the explanation of the tricks extends to parts of temporal and parietal cortex, whereas significant decoding after revelation is largely restricted to posterior visual areas. In sum, and in contrast to the univariate results that showed no net modulation as a function of prior knowledge (and surprise) in visual areas, the differences in decoding accuracies using data before and after explanation of the tricks suggests that visual areas are indeed sensitive to changes in knowledge and encode specific expectations.

Controlling for time effects in pattern activity

Arguably, suprathreshold decoding of magic effects in posterior sensory areas may reflect general stimulus differences in the moment of magic between the different magic trick types independent of the object used, as appear videos systematically showed red objects, color changing videos blue objects and vanishing videos no object. However, the observed significant differences in decoding accuracy before and after revelation suggest surprise-dependent modulations of activity patterns, as these differences are present in view of the very same videos. Yet, these differences could reflect time- and/or design-related confounds, such as a general decrease of attention and alertness over time. However, since we did not find any significant changes in univariate comparisons in posterior visual areas and we observed a general increase in parts of the dorsal attention network, we believe that our results are not confounded by time and/or design related factors. Nonetheless we performed control analyses comparing decoding accuracy of objects present or absent in control videos in the pre vs post revelation phase. These analyses showed no modulation of time in control videos (detailed results can be found in the Supplementary Materials Section *MVPA control analyses)*.

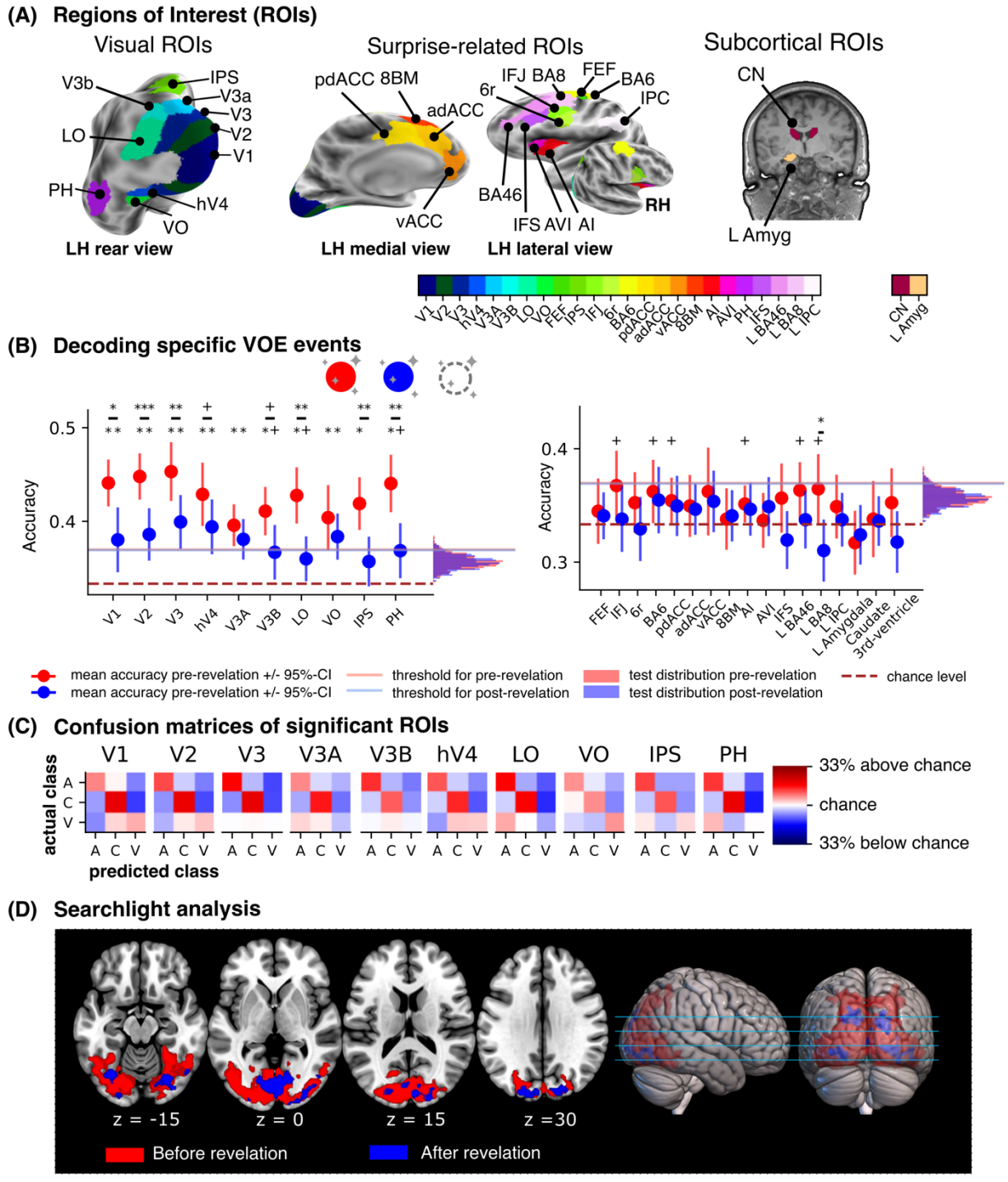

(A) Regions of Interest (ROIs)
Visual ROIs
Surprise-related ROIs
Subcortical ROIs
V3b
IPS
V3a
V3
V2
V1
LO
PH
hV4
VO
LH rear view
pdACC 8BM
adACC
vACC
LH medial view
IFJ BA8
FEF
BA6
6r
IPC
BA46
IFS AVI AI
RH
LH lateral view
CN
L Amyg
v1 V2 V3 hV4 V3A V3B LO VO FEF IPS IFJ 6r BA6 pdACC adACC vACC 8BM AI AVI PH IFS L BA46 L BA8 L IPC
CN L Amyg
(B) Decoding specific VOE events
Accuracy
0.5
0.4
V1 V2 V3 hV4 V3A V3B LO VO IPS PH
0.4
0.3
FEF IFJ 6r BA6 pdACC adACC vACC 8BM AI AVI IFS L BA46 L BA8 L IPC L Amygdala Caudate 3rd-ventricle
mean accuracy pre-revelation +/- 95%-CI
mean accuracy pre-revelation +/- 95%-CI
threshold for pre-revelation
threshold for post-revelation
test distribution pre-revelation
test distribution post-revelation
chance level
(C) Confusion matrices of significant ROIs
V1 V2 V3 V3A V3B hV4 LO VO IPS PH
actual class
A C V
predicted class
33% above chance
chance
33% below chance
(D) Searchlight analysis
z = -15
z = 0
z = 15
z =30
Before revelation
After revelation

Figure 5. Results of decoding analyses **A**, regions of interest (ROIs) used in the experiment. Visual ROIs are shown left, all of which (except for the inferior temporal area PH) were defined using a probabilistic map of visual areas (Wang et al., 2015). Surprise-related ROIs presented on the right panel were defined based on previous results (Danek et al., 2015; Parris et al., 2009) using a multimodal parcellation atlas (Glasser et al., 2016), except for frontal eye field (FEF), which was also defined using the probabilistic atlas from Wang et al., (2015) (see supplementary section *Surprise-related region of Interest definition* for more information). The two subcortical ROIs, caudate nucleus, and left amygdala were defined using the Freesurfer automatic parcellation (Fischl et al., 2002). **B**, shown are the decoding accuracies for decoding the VOE types over objects in a three-fold cross-decoding approach in our theory-driven ROIs (left: ROIs that significantly decoded the VOE type using pre-revelation data, right: ROIs that did *not* significantly decode the VOE type using pre-revelation data). Decoding was performed with data before (red) and after (blue) revelation. All statistics were corrected for multiple tests by using the max-statistic correction across all ROIs (Nichols and Holmes, 2002) **C**, confusion matrices from decoding VOE types before revelation for ROIs that could significantly decode VOE types before revelation. It appears that VOEs due to unexpected color changes are predicted best, while the VOEs due to objects vanishing is predicted less than VOEs due to objects appearing and changing. **D**, whole-brain searchlight decoding results. We can significantly decode VOE types in the majority of visual cortex before revelation and less so after revelation (correcting using a permutation-bootstrap hybrid method, Stelzer et al., 2013).

Discussion

In this fMRI study, we used naturalistic video stimuli showing magic tricks and matched control actions to investigate responses to violation of expectations (VOE) of deeply held beliefs about the physical world. We used three distinct magic types (object appearance, object disappearance, and feature-change) that were presented with and without prior knowledge about the underlying deceptive methods (i.e., sleights-of-hand). Each magic type was presented using three distinct objects to allow for object-invariant classification of magic types. We looked for 1) generic prediction error responses to perceived violation of physical principles, 2) specific responses to the different magic types and, 3) effects of the viewers' prior knowledge on prediction error processing, for both generic and specific responses.

Our results revealed a hierarchy of surprise signals. First, we observed generic effects of world-model VOE (i.e. common to all magic types) in several clusters of the prefrontal and parietal cortex (such as the dorsal and ventral ACC and the posterior parietal cortex). Then, differential activity specific to the different types of magic was evoked predominately in posterior visual areas of the occipital and parietal cortex. These specific prediction error signals were evident in the univariate analyses and in decoding of the magic types, both of which were confined to posterior areas across the visual hierarchy. Finally, following explanation of the tricks, responses were largely unaffected by participants knowledge and only decreased in select parts of the network showing generic effects of VOE (midline areas of the default mode network). While net activity in visual areas was not significantly modulated by the prior knowledge, decoding of VOE type-specific signals was sensitive to changes in the participants knowledge, showing decreased decoding when participants knew the tricks. These results suggest that higher-level predictive information affects even the earliest levels of cortical visual processing (V1-V3).

Generic responses to violation of expectations

Witnessing magic events that violate intuitive physical principles evoked activity in a large network of frontoparietal (dACC, vACC/mPFC and posterior parietal cortex) and subcortical (caudate nucleus) areas, with no involvement of lower-level sensory areas. This pattern of activity is consistent with that observed in a recent large meta-analysis of surprising events (Fouragnan et al., 2018) and with previous experiments investigating surprise responses using naturalistic videos, such as magic tricks (Danek et al., 2015), computer generated animations (Bardi et al., 2017) or learned sequences of movements (Schiffer and Schubotz, 2011). Accordingly, our results add to prior evidence showing the key role of the dACC in processing incongruent information (Alexander and Brown, 2019, 2011) and of the caudate nucleus in signaling unexpected and rewarding events (Schultz et al., 1997; Wittmann et al., 2008; Zink et al., 2003). Most importantly, our results suggest that higher-level VOE in view of seemingly impossible events are processed similarly to breaches of lower-level expectations, such as the presence of infrequent stimuli or

unlikely events (cf. Grassi and Bartels, 2021). This suggests the existence of a dedicated frontoparietal network signaling the detection of incongruent information in the human brain.

Specific responses to violation of expectations

Complementing the generic frontoparietal involvement in processing naturalistic violations of physical principles, we further looked into the specific effects of the different types of VOE (appear, change, vanish). Specific responses evoked by the different VOE in net activity and multivariate activation patterns (i.e., allowing for a distinction between trick-types) were observed exclusively in posterior sensory areas. In contrast, frontal areas revealed no or only weak specific VOE responses. Since this divergent pattern of net activity was only observable when looking into the individual VOE, it is possible that previous studies failed to report the involvement of sensory areas because of pooling responses to different types of VOE (Danek et al., 2015; Liu et al., 2024; Parris et al., 2009).

Previous results using dynamically occluded stimuli report the neural representation of occluded objects in posterior visual areas (Erlikhman and Caplovitz, 2017; Hulme and Zeki, 2007; Olson et al., 2004) and in neurons of the inferotemporal cortex of macaque monkeys (Puneeth and Arun, 2016) at different levels of complexity (e.g., occluded faces selectively engaged the fusiform face areas, Hulme and Zeki, 2007). The observed differential activity in visual areas using naturalistic stimuli in the present study are likely VOE type-specific surprise signals when violating said representations.

The following reasons support this interpretation: First, net responses were not driven by differences in visual content (because they were evident across distinct objects, and additional control contrasts ruled content-driven responses out). Second, decoding did not work in the absence of VOE when we used similar sensory occurrences using data from the matched control videos. Third, prior knowledge significantly reduced decoding accuracies. And finally, the observed responses occurred in functionally specialized regions of the visual cortex. For example, the unexpected appearance of objects evoked activity in the object-responsive LOC and the perception of unexpected colors evoked activity in the ventral color areas of the fusiform gyrus. Both of which are compatible with prior evidence showing increased activity in the LOC upon perceiving unexpected objects (Richter et al., 2018) and in color areas when viewing unexpected colors (Jiang et al., 2016; Stefanics et al., 2019). Hence, our results suggest that memory-based expectations related to higher-level principles affect visual processing already at the earliest cortical visual processing areas: they encode information about the presence, absence, and features of objects.

Please note that the specific responses observed may additionally be related to attentional mechanisms engaged in processing the feature-specific VOE. Indeed, attentional mechanisms are likely entangled with processing of VOE, to enhance model updating (Hohwy, 2012). However, differences in attention are unlikely the sole account of our VOE-specific results, as we observed

an enhanced involvement of the dorsal attention network (DAN) post-explanation compared to pre-explanation, while the VOE-type specific activity remained constant.

Hierarchical prediction errors in naturalistic perception

The observation of surprise-related information in posterior visual areas is in line with a variety of higher-level memory-based signals that have been reported in visual areas, such as memory color (Bannert and Bartels, 2013), scene context (Muckli et al., 2015), scene segmentation (Grassi et al., 2018, 2017; Scholte et al., 2008), expected visual stimuli (Ekman et al., 2017) and working memory (Harrison and Tong, 2009). Importantly, these signals have been interpreted as evidence of recurrent predictive signals from higher-level areas, as they encoded information that is not thought to originate from V1 (such as memory color, 3D or Gestalt and scene information). Consistent with this, further studies located corresponding signals in superficial and/or deeper layers of the cortex using laminar fMRI (Aitken et al., 2020; Lawrence et al., 2018; Muckli et al., 2015) or electrophysiological measurements in monkeys (e.g., Papale et al., 2022; Self et al., 2013).

Together, the current results fall in line with predictive coding theories and extend them to VOE regarding higher-level world models (Friston, 2005; Lee and Mumford, 2003; Rao and Ballard, 1999). We show a clear dissociation: generic responses to VOE in higher-level frontoparietal areas and segregated surprise responses in functionally specialized lower-level sensory areas (involved in the processing of the expected information). This reflects the hierarchical structure of our internal world model (Clark, 2013; Hohwy, 2014), with frontoparietal areas involved in representing more abstract aspects of the world (such as object permanency), while sensory areas represent lower-level inferences about the immediate and detailed features (such as color and shapes). Accordingly, the observed surprise-related responses in lower-level sensory areas can be thought of the product of a mismatch between top-down predictions (“a red ball”) fed back to lower-level areas to be compared with incoming sensory evidence (“a blue ball”) (Grassi and Bartels, 2021).

Knowledge-dependent modulations

To further investigate these hierarchical VOE responses, we probed how prior knowledge affected them. We hypothesized that providing the participants with knowledge about the mechanics of the trick for each video would avert VOE: why should we be surprised when viewing a disappearing ball when we know how and that the magician is actually hiding it behind his hands?

Surprisingly, while participants subjective surprise ratings were significantly reduced after providing them with the explanation of the tricks, net brain responses were almost indistinguishable: areas involved in the processing of unexpected events, such as the dACC, anterior insula and caudate nucleus (Fouragnan et al., 2018) were systematically active when

observing the magic videos even after participants had rational explanations for the tricks. As neural responses were reminiscent to those signaling surprise, it suggests that repeated viewing of explainable events did not prevent VOE. This intriguing observation likely reflects that people can be moved by things they know to be unreal, such as fictions (i.e., the "paradox of fiction", see Radford and Weston, 1975) or magic illusions (i.e., the "paradox of theatrical magic", see Grassi et al., 2024). This is akin to how we still perceive visual illusions, even if we know how they work. For example, when a magician convincingly saws someone in half on stage, the audience is genuinely moved by the illusion (i.e., surprised), but do not attempt to prevent it nor call the police (because they know it is unreal). Our results suggest that the compelling perceptual illusions that magic provides are initially appraised as surprising, even with existing prior-information (cf. Grassi et al., 2024).

In turn, the only areas whose activity decreased following the explanation of the tricks were two midline core areas of the default mode network (DMN), the ventral ACC/mPFC and the posterior cingulate cortex. The modulation of midline core areas of the DMN by prior knowledge is consistent with recent reports showing their involvement in processing surprising events in movies (Brandman et al., 2021), jokes (Jääskeläinen et al., 2016) and structured events unfolding in time (Baldassano et al., 2018; Regev et al., 2013; Simony et al., 2016). Based on these findings, it has been suggested that the DMN is not to be understood exclusively as an “intrinsic” network (as originally proposed, cf. Raichle, 2015), but as a dynamic “sense-making” network involved in the creation of rich models of events by integrating incoming information with prior knowledge as they unfold over time instead (Stawarczyk et al., 2021; Yeshurun et al., 2021).

Here, we show that areas of the “sense-making” network may be sensitive to the rational explanation of magic tricks, whereas all other identified surprise-related regions continued to be sensitive to the VOE (even once the tricks were understood). The decreased involvement of areas of the DMN, together with an increase of activity in frontoparietal areas of the dorsal attention network (DAN), is consistent with a reduction in prediction error (surprise) signals related to narrative understanding (engaging the DMN) and an increase of top-down attention after explanation of the tricks (engaging the DAN).

Conclusion

We used a naturalistic paradigm to violate deeply held beliefs of our physical world, involving three types of expectation violations (object appearance, color change, and object disappearance). Our results show a hierarchy of surprise signals: generic responses to unexpected events in frontoparietal areas, and responses specific to the type of VOE in distinct functionally specialized sensory areas. Our results suggest that world-model VOE are processed similarly to other surprising events in dedicated areas of the prefrontal cortex and striatum, and that core midline areas of the default-mode network decrease their involvement once rational understanding is

established. Most importantly, we show that early and functionally specialized areas of the visual cortex encode memory-based predictions about the presence, absence, and features of objects.

Data accessibility

Code (for preprocessing, analysis and visualisations) and preprocessed data for group analyses can be found at https://osf.io/kn2af/?view_only=067b698a4567441e93b01518a88860a0. Raw MRI data can be provided upon reasonable request.

CRediT author statement

Vincent Plikat: conceptualization, formal analysis (lead), investigation (equal), visualisation, writing – original draft, writing – review and editing. **Pablo R. Grassi:** conceptualization (lead), formal analysis (supporting), investigation (equal), methodology, supervision (equal), visualization, writing – original draft, writing – review and editing. **Julius Frack:** resources. **Andreas Bartels:** conceptualization, methodology, supervision (equal), writing – review and editing.

Acknowledgements

This work was supported by Barbara-Wengeler-Foundation, the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) project number 465409366, and by the Max Planck Society.

References

Abraham A, Pedregosa F, Eickenberg M, Gervais P, Mueller A, Kossaifi J, Gramfort A, Thirion B, Varoquaux G. 2014. Machine learning for neuroimaging with scikit-learn. *Front Neuroinform* **8**. doi:10.3389/fninf.2014.00014

Adams WJ, Graf EW, Ernst MO. 2004. Experience can change the “light-from-above” prior. *Nature Neuroscience* **7**:1057–1058. doi:10.1038/nn1312

Aitken F, Menelaou G, Warrington O, Koolschijn RS, Corbin N, Callaghan MF, Kok P. 2020. Prior expectations evoke stimulus-specific activity in the deep layers of the primary visual cortex. *PLoS Biol* **18**:e3001023. doi:10.1371/journal.pbio.3001023

Alexander WH, Brown JW. 2019. The Role of the Anterior Cingulate Cortex in Prediction Error and Signaling Surprise. *Topics in Cognitive Science* **11**:119–135. doi:10.1111/tops.12307

Alexander WH, Brown JW. 2011. Medial prefrontal cortex as an action-outcome predictor. *Nature Neuroscience* **14**:1338–1344. doi:10.1038/nn.2921

Baldassano C, Hasson U, Norman KA. 2018. Representation of Real-World Event Schemas during Narrative Perception. *J Neurosci* **38**:9689–9699. doi:10.1523/JNEUROSCI.0251-18.2018

Bannert MM, Bartels A. 2013. Decoding the yellow of a gray banana. *Current Biology* **23**:2268–2272. doi:10.1016/j.cub.2013.09.016

Bardi L, Desmet C, Nijhof A, Wiersema JR, Brass M. 2017. Brain activation for spontaneous and explicit false belief tasks overlaps: New fMRI evidence on belief processing and violation of expectation. *Social Cognitive and Affective Neuroscience* **12**:391–400. doi:10.1093/scan/nsw143

Bartels A, Zeki S. 2000. The architecture of the colour centre in the human visual brain: New results and a review. *European Journal of Neuroscience* **12**:172–193. doi:10.1046/j.1460-9568.2000.00905.x

Battaglia PW, Hamrick JB, Tenenbaum JB. 2013. Simulation as an engine of physical scene understanding. *Proc Natl Acad Sci USA* **110**:18327–18332. doi:10.1073/pnas.1306572110

Brandman T, Malach R, Simony E. 2021. The surprising role of the default mode network in naturalistic perception. *Commun Biol* **4**:79. doi:10.1038/s42003-020-01602-z

Clark A. 2013. Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences* **36**:181–204. doi:10.1017/S0140525X12000477

Corbetta M, Shulman GL. 2002. Control of goal-directed and stimulus-driven attention in the brain. *Nature reviews Neuroscience* **3**:201–15. doi:10.1038/nrn755

Dale AM, Fischl B, Sereno MI. 1999. Cortical Surface-Based Analysis. *NeuroImage* **9**:179–194. doi:10.1006/nimg.1998.0395

Danek AH, Öllinger M, Fraps T, Grothe B, Flanagin VL. 2015. An fMRI investigation of expectation violation in magic tricks. *Frontiers in Psychology* **6**:1–11. doi:10.3389/fpsyg.2015.00084

de Lange FP, Heilbron M, Kok P. 2018. How Do Expectations Shape Perception? *Trends in Cognitive Sciences* **22**:764–779. doi:10.1016/j.tics.2018.06.002

Egner T, Monti JM, Summerfield C. 2010. Expectation and surprise determine neural population responses in the ventral visual stream. *Journal of Neuroscience* **30**:16601–16608. doi:10.1523/JNEUROSCI.2770-10.2010

Ekman M, Kok P, De Lange FP. 2017. Time-compressed preplay of anticipated events in human primary visual cortex. *Nature Communications* **8**:1–9. doi:10.1038/ncomms15276

Erlikhman G, Caplovitz GP. 2017. Decoding information about dynamically occluded objects in visual cortex. *NeuroImage* **146**:778–788. doi:10.1016/j.neuroimage.2016.09.024

Fischer J, Mikhael JG, Tenenbaum JB, Kanwisher N. 2016. Functional neuroanatomy of intuitive physical inference. *Proc Natl Acad Sci USA* **113**. doi:10.1073/pnas.1610344113

Fischl B, Salat DH, Busa E, Albert M, Dieterich M, Haselgrove C, van der Kouwe A, Killiany R, Kennedy D, Klaveness S, Montillo A, Makris N, Rosen B, Dale AM. 2002. Whole Brain Segmentation. *Neuron* **33**:341–355. doi:10.1016/S0896-6273(02)00569-X

Fouragnan E, Retzler C, Philiastides MG. 2018. Separate neural representations of prediction error valence and surprise: Evidence from an fMRI meta-analysis. *Hum Brain Mapp* **39**:2887–2906. doi:10.1002/hbm.24047

Friston K. 2010. The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience* **11**:127–138. doi:10.1038/nrn2787

Friston K. 2005. A theory of cortical responses. *Philosophical transactions of the Royal Society of London Series B, Biological sciences* **360**:815–36. doi:10.1098/rstb.2005.1622

Friston KJ, Penny WD, Glaser DE. 2005. Conjunction revisited. *NeuroImage* **25**:661–7. doi:10.1016/j.neuroimage.2005.01.013

Glasser MF, Coalson TS, Robinson EC, Hacker CD, Harwell J, Yacoub E, Ugurbil K, Andersson J, Beckmann CF, Jenkinson M, Smith SM, Van Essen DC. 2016. A multi-modal parcellation of human cerebral cortex. *Nature* **536**:171–178. doi:10.1038/nature18933

Grassi PR, Bartels A. 2021. Magic, Bayes and wows: A Bayesian account of magic tricks. *Neuroscience & Biobehavioral Reviews* **126**:515–527. doi:10.1016/j.neubiorev.2021.04.001

Grassi PR, Plikat V, Wong HY. 2024. How can we be moved by magic? *British Journal of Aesthetics* **64**:187–204. doi:10.1093/aesthj/ayad026

Grassi PR, Zaretskaya N, Bartels A. 2018. A Generic Mechanism for Perceptual Organization in the Parietal Cortex. *The Journal of Neuroscience* **38**:7158–7169. doi:10.1523/JNEUROSCI.0436-18.2018

Grassi PR, Zaretskaya N, Bartels A. 2017. Scene segmentation in early visual cortex during suppression of ventral stream regions. *NeuroImage* **146**:71–80. doi:10.1016/j.neuroimage.2016.11.024

Grill-Spector K, Kourtzi Z, Kanwisher N. 2001. The lateral occipital complex and its role in object recognition. *Vision Research* **41**:1409–1422. doi:10.1016/S0042-6989(01)00073-6

Harrison SA, Tong F. 2009. Decoding reveals the contents of visual working memory in early visual areas. *Nature* **458**:632–635. doi:10.1038/nature07832

Hespos SJ, Ferry AL, Rips LJ. 2009. Five-Month-Old Infants Have Different Expectations for Solids and Liquids. *Psychol Sci* **20**:603–611. doi:10.1111/j.1467-9280.2009.02331.x

Hohwy J. 2014. The Predictive Mind. Oxford: OUP Oxford. doi:10.1093/acprof:oso/9780199682737.001.0001

Hohwy J. 2012. Attention and conscious perception in the hypothesis testing brain. *Frontiers in Psychology* **3**:1–14. doi:10.3389/fpsyg.2012.00096

Hohwy J, Roepstorff A, Friston K. 2008. Predictive coding explains binocular rivalry: an epistemological review. *Cognition* **108**:687–701. doi:10.1016/j.cognition.2008.05.010

Hulme OJ, Zeki S. 2007. The Sightless View: Neural Correlates of Occluded Objects. *Cerebral Cortex* **17**:1197–1205. doi:10.1093/cercor/bhl031

Jääskeläinen IP, Pajula J, Tohka J, Lee H-J, Kuo W-J, Lin F-H. 2016. Brain hemodynamic activity during viewing and re-viewing of comedy movies explained by experienced humor. *Sci Rep* **6**:27741. doi:10.1038/srep27741

Jiang J, Summerfield C, Egner T. 2016. Visual Prediction Error Spreads Across Object Features in Human Visual Cortex. *J Neurosci* **36**:12746–12763. doi:10.1523/JNEUROSCI.1546-16.2016

Kok P, Jehee JFM, de Lange FP. 2012a. Less Is More: Expectation Sharpens Representations in the Primary Visual Cortex. *Neuron* **75**:265–270. doi:10.1016/j.neuron.2012.04.034

Kok P, Rahnev D, Jehee JFM, Lau HC, de Lange FP. 2012b. Attention Reverses the Effect of Prediction in Silencing Sensory Signals. *Cerebral Cortex* **22**:2197–2206. doi:10.1093/cercor/bhr310

Lawrence SJD, van Mourik T, Kok P, Koopmans PJ, Norris DG, de Lange FP. 2018. Laminar Organization of Working Memory Signals in Human Visual Cortex. *Current Biology*. doi:10.1016/j.cub.2018.08.043

Lee TS, Mumford D. 2003. Hierarchical Bayesian inference in the visual cortex. *Journal of the Optical Society of America A* **20**:1434. doi:10.1364/josaa.20.001434

Liu S, Lydic K, Mei L, Saxe R. 2024. Violations of physical and psychological expectations in the human adult brain. *Imaging Neuroscience* **2**:1–25. doi:10.1162/imag_a_00068

Muckli L, De Martino F, Vizioli L, Petro LS, Smith FW, Ugurbil K, Goebel R, Yacoub E. 2015. Contextual Feedback to Superficial Layers of V1. *Current Biology* **25**:2690–2695. doi:10.1016/j.cub.2015.08.057

Nichols TE, Holmes AP. 2002. Nonparametric permutation tests for functional neuroimaging: A primer with examples. *Human Brain Mapping* **15**:1–25. doi:10.1002/hbm.1058

Notter M, Gale D, Herholz P, Markello R, Notter-Bielser M-L, Whitaker K. 2019. AtlasReader: A Python package to generate coordinate tables, region labels, and informative figures from statistical MRI images. *JOSS* **4**:1257. doi:10.21105/joss.01257

Olson IR, Gatenby JC, Leung H-C, Skudlarski P, Gore JC. 2004. Neuronal representation of occluded objects in the human brain. *Neuropsychologia* **42**:95–104. doi:10.1016/S0028-3932(03)00151-9

Papale P, Wang F, Morgan AT, Chen X, Gilhuis A, Petro LS, Muckli L, Roelfsema PR, Self MW. 2022. Feedback brings scene information to the representation of occluded image regions in area V1 of monkeys and humans. doi:10.1101/2022.11.21.517305

Parris BA, Kuhn G, Mizon GA, Benattayallah A, Hodgson TL. 2009. Imaging the impossible: An fMRI study of impossible causal relationships in magic tricks. *NeuroImage* **45**:1033–1039. doi:10.1016/j.neuroimage.2008.12.036

Pedregosa F. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**:2825–2830.

Puneeth NC, Arun SP. 2016. A neural substrate for object permanence in monkey inferotemporal cortex. *Sci Rep* **6**:30808. doi:10.1038/srep30808

Radford C, Weston M. 1975. How Can We Be Moved by the Fate of Anna Karenina? *Aristot Soc Suppl Vol* **49**:67–94. doi:10.1093/aristoteliansupp/49.1.67

Raichle ME. 2015. The Brain's Default Mode Network. *Annu Rev Neurosci* **38**:433–447. doi:10.1146/annurev-neuro-071013-014030

Rao RPN, Ballard DH. 1999. Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience* **2**:79–87. doi:10.1038/4580

Regev M, Honey CJ, Simony E, Hasson U. 2013. Selective and Invariant Neural Responses to Spoken and Written Narratives. *Journal of Neuroscience* **33**:15978–15988. doi:10.1523/JNEUROSCI.1580-13.2013

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

Rolls ET, Huang C-C, Lin C-P, Feng J, Joliot M. 2020. Automated anatomical labelling atlas 3. *NeuroImage* **206**:116189. doi:10.1016/j.neuroimage.2019.116189

SanMiguel I, Widmann A, Bendixen A, Trujillo-Barreto N, Schroger E. 2013. Hearing Silences: Human Auditory Processing Relies on Preactivation of Sound-Specific Brain Activity Patterns. *Journal of Neuroscience* **33**:8633–8639. doi:10.1523/JNEUROSCI.5821-12.2013

Schiffer A-M, Schubotz RI. 2011. Caudate Nucleus Signals for Breaches of Expectation in a Movement Observation Paradigm. *Front Hum Neurosci* **5**. doi:10.3389/fnhum.2011.00038

Scholte HS, Jolij J, Fahrenfort JJ, Lamme VAF. 2008. Feedforward and recurrent processing in scene segmentation: Electroencephalography and functional magnetic resonance imaging. *Journal of Cognitive Neuroscience* **20**:2097–2109. doi:10.1162/jocn.2008.20142

Schultz W, Dayan P, Montague PR. 1997. A neural substrate of prediction and reward. *Science* **275**:1593–1599. doi:10.1126/science.275.5306.1593

Schwettmann S, Tenenbaum JB, Kanwisher N. 2019. Invariant representations of mass in the human brain. *eLife* **8**:e46619. doi:10.7554/eLife.46619

Self MW, van Kerkoerle T, Supèr H, Roelfsema PR. 2013. Distinct Roles of the Cortical Layers of Area V1 in Figure-Ground Segregation. *Current Biology* **23**:2121–2129. doi:10.1016/j.cub.2013.09.013

Simony E, Honey CJ, Chen J, Lositsky O, Yeshurun Y, Wiesel A, Hasson U. 2016. Dynamic reconfiguration of the default mode network during narrative comprehension. *Nat Commun* **7**:12141. doi:10.1038/ncomms12141

Stawarczyk D, Bezdek MA, Zacks JM. 2021. Event Representations and Predictive Processing: The Role of the Midline Default Network Core. *Top Cogn Sci* **13**:164–186. doi:10.1111/tops.12450

Stefanics G, Stephan KE, Heinzle J. 2019. Feature-specific prediction errors for visual mismatch. *NeuroImage* **196**:142–151. doi:10.1016/j.neuroimage.2019.04.020

Stelzer J, Chen Y, Turner R. 2013. Statistical inference and multiple testing correction in classification-based multi-voxel pattern analysis (MVPA): Random permutations and cluster size control. *NeuroImage* **65**:69–82. doi:10.1016/j.neuroimage.2012.09.063

Todorovic A, van Ede F, Maris E, de Lange FP. 2011. Prior expectation mediates neural adaptation to repeated sounds in the auditory cortex: An MEG study. *Journal of Neuroscience* **31**:9118–9123. doi:10.1523/JNEUROSCI.1425-11.2011

Vizioli L, Moeller S, Dowdle L, Akçakaya M, De Martino F, Yacoub E, Uğurbil K. 2021. Lowering the thermal noise barrier in functional brain mapping with magnetic resonance imaging. *Nat Commun* **12**:5181. doi:10.1038/s41467-021-25431-8

Wacongne C, Labyt E, Van Wassenhove V, Bekinschtein T, Naccache L, Dehaene S. 2011. Evidence for a hierarchy of predictions and prediction errors in human cortex. *Proceedings of the National Academy of Sciences of the United States of America* **108**:20754–20759. doi:10.1073/pnas.1117807108

Wang L, Mruczek REB, Arcaro MJ, Kastner S. 2015. Probabilistic maps of visual topography in human cortex. *Cerebral Cortex* **25**:3911–3931. doi:10.1093/cercor/bhu277

Wang S. 2004. Young infants' reasoning about hidden objects: evidence from violation-of-expectation tasks with test trials only. *Cognition* **93**:167–198. doi:10.1016/j.cognition.2003.09.012

Wessel JR, Danielmeier C, Morton JB, Ullsperger M. 2012. Surprise and Error: Common Neuronal Architecture for the Processing of Errors and Novelty. *Journal of Neuroscience* **32**:7528–7537. doi:10.1523/JNEUROSCI.6352-11.2012

Wittmann BC, Daw ND, Seymour B, Dolan RJ. 2008. Striatal Activity Underlies Novelty-Based Choice in Humans. *Neuron* **58**:967–973. doi:10.1016/j.neuron.2008.04.027

Wynn K. 1992. Addition and subtraction by human infants. *Nature* **358**:749–750. doi:10.1038/358749a0

Yeo BTT, Krienen FM, Sepulcre J, Sabuncu MR, Lashkari D, Hollinshead M, Roffman JL, Smoller JW, Zollei L, Polimeni JR, Fischl B, Liu H, Buckner RL. 2011. The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *Journal of neurophysiology* **106**:1125–1165. doi:10.1152/jn.00338.2011.

Yeshurun Y, Nguyen M, Hasson U. 2021. The default mode network: where the idiosyncratic self meets the shared social world. *Nat Rev Neurosci* **22**:181–192. doi:10.1038/s41583-020-00420-w

Zink CF, Pagnoni G, Martin ME, Dhamala M, Berns GS. 2003. Human Striatal Response to Salient Nonrewarding Stimuli. *J Neurosci* **23**:8092–8097. doi:10.1523/JNEUROSCI.23-22-08092.2003