Violations of physical and psychological expectations in the human adult brain

Shari Liu [1], Kirsten Lydic [2], Lingjie Mei [3], & Rebecca Saxe [4]

[1] Dept Psychological and Brain Sciences, Johns Hopkins
[2] Annenberg School for Communication, University of Pennsylvania
[3] Dept Computer Science, Princeton University
[4] Dept Brain and Cognitive Sciences, Massachusetts Institute of Technology

Author note: Correspondence concerning this article should be addressed to Shari Liu (shariliu@jhu.edu) or Rebecca Saxe (saxe@mit.edu).

Abstract

After seeing one solid object apparently passing through another, or a person taking the long route to a destination when a shortcut was available, human adults classify those events as surprising. When tested on these events in violation-of-expectation (VOE) experiments, infants look longer at the same outcomes, relative to similar but expected outcomes. What cognitive processes underlie these judgments from adults, and perhaps infants' sustained attention to these events? As one approach to test this question, we used functional magnetic resonance imaging (fMRI) to scan the brains of human adults (total N = 49, 22 female, mean age of 26 years) while they viewed stimuli that were originally designed to test for physical and psychological expectations in infants. We examined non-mutually exclusive candidates for the processes underlying the VOE effect, including domain-general processes, like visual prediction error and curiosity, and domain-specific processes, like prediction error with respect to distinctively physical and psychological expectations (objects are solid; agents behave rationally). Early visual regions did not distinguish between expected and unexpected events from either domain. By contrast, multiple demand regions, involved in goal-directed attention, responded more to unexpected events in both domains, providing evidence for domain-general goal-directed attention as a mechanism for VOE. Left supramarginal gyrus (LSMG) was engaged during physical prediction and responded preferentially to unexpected events from the physical domain, providing evidence for domain-specific physical prediction error. Thus, in adult brains, violations of physical and psychological expectations involve domain-specific, and domain-general, though not purely visual, computations.

Significance Statement

When an object hovers in midair, or a person acts irrationally, infants look and pay attention to those events. What mental processes account for this behavior: that these events are visually novel, evoke curiosity, and/or violate infants' expectations about the physical and psychological world? We scanned adults using functional magnetic resonance imaging and found that adults do not merely process such events as novel visual stimuli. Instead, these events evoke distinctively physical and psychological processing, as well as domain-general, internally driven attention. These results serve as a baseline for future studies of infants and illustrate the promise of using the tools of cognitive neuroscience to address questions about infant minds.

Main Text

In the first year of life, human infants rapidly develop expectations about the properties and behavior of inanimate objects, and animate agents. Like adults, they distinguish between surprising events and visually similar but unsurprising events (e.g. a ball rolls off the edge of a table, and hovers in midair, or stops rolling before it reaches the edge of the table). Infants look longer at the unexpected than expected outcome (the violation-of-expectation, or VOE, response) towards many events that adults rate as surprising (1, 2): for example, when objects float in midair (3) or appear to pass through each other (4), and when agents change their goals (5) or act inefficiently (6). The mental processes that drive longer looking in these studies are still hotly debated (7, 8). Do infants respond to these events in virtue of domain-specific expectations about psychological and physical events (9, 10)? Or are there stimulus-driven alternative explanations that could also explain these patterns of behavior (11, 12)? And do longer looking in infants, and judgments of surprise in adults, reflect the detection of a surprising outcome, or also motivation to explore and explain the source of surprise (13, 14)?

Domain-specific hypotheses

One hypothesis regarding VOE effects in the developmental psychology literature is that surprising events violate *distinctively physical and psychological expectations*: that objects are solid and permanent; and that agents act efficiently towards goals. The strongest version of this hypothesis is that infants possess ‘core knowledge': an early-emerging conceptual repertoire consisting of domain-specific systems for different domains of thought, including physics, psychology, number, and space (15). There is evidence from developmental psychology that infants have distinct expectations for agents and objects: Infants represent objects as solid and permanent entities that do not hover in midair, or blip in and out of existence (10). Infants represent agents as actors who have goals, and pursue them in consistent and efficient ways (6). There is also evidence that infants have some shared expectations across both domains. For example, infants expect that both agents and objects are solid entities (16).

Domain-general hypotheses

Another broad hypothesis under consideration is that surprising events from violation-of-expectation studies evoke domain-general processes. One such process is *stimulus-driven prediction error* (i.e. a response to the visual features of the unexpected stimulus). While infant looking-time studies typically account for some simple perceptual alternative explanations, infants do reliably look longer at scenes that are visually novel (17, 18). Furthermore, unexpected and expected events must be visually distinguishable, and thus each pair of events differs along at least one visual dimension. Developmental psychologists remain divided about whether for any pair of VOE stimuli, longer looking may be driven by distinctive visual features (19, 20).

A second domain-general hypothesis is that unexpected physical and psychological events evoke *curiosity and motivation* to gain information about the source of surprise (13, 14). Under this hypothesis, infant looking is not merely a passive behavior, but rather an active process driven by the infant’s own learning goals (21). There is some evidence that unexpected events evoke curiosity in infants. After viewing an unexpected physical event, such as a ball rolling through a solid wall, infants show enhanced learning about that object (22), and choose to explore that object (23) as though they are trying to explain the outcome (e.g. by banging the ball after seeing a violation of solidity, and dropping the ball after seeing a violation of support) (22). In addition, the VOE response only arises when infants have reason to be curious: Infants look longer when a ball passes through a solid wall, rather than stopping short of the wall, but not if they first see that the wall has an archway through it, allowing the ball to pass through (24).

The contribution of functional neuroimaging to testing these hypotheses
Plausibly, all of these mental processes could influence infant looking, but which of these accounts for the VOE response? Despite decades of behavioral work, controversy remains. Here, we consider the potential contribution of neuroimaging to this debate, which can reveal the hidden internal processes underlying VOE by studying them simultaneously and directly.

If domain-specific processing underlies the VOE response, what brain regions could support those computations? In adults, different cortical regions represent the properties and dynamics of agents and objects. A set of regions including the temporoparietal junction (TPJ), medial prefrontal cortex (MPFC), precuneus (PC), and superior temporal sulcus (STS) are engaged during social perception and cognition (25, 26). The STS, in particular, tracks other people's actions, intentions, and interactions (27–32). A distinct set of regions including supplementary motor area, superior parietal cortex, and supramarginal gyrus (SMG), represents physical information including object mass and stability (33–35). As early as has been measured, similar regions in infants are implicated in the processing of social vs physical stimuli (36–39), making studying these regions in adults relevant to hypotheses about the minds and brains of infants. Prior work measuring neural responses towards surprising physical and psychological stimuli has reported increased neural activity toward unexpected outcomes in regions associated with social processing, as well as domain-general multiple demand (40–43), consistent with a neural prediction error (44): an increased response that encodes the difference between what was expected and what was observed. If these regions compute domain-specific prediction error in VOE events, then we expect to observe greater activity in each of these regions for unexpected events from the matching domain (e.g. a greater response to unexpected than expected physical events in SMG, and to unexpected than expected psychological events in STS).

By contrast, if early visual processing underlies the VOE response, then which regions would we expect to support this process? Early visual regions, including the primary visual cortex (V1) and motion-sensitive area (MT), are sensitive to a host of low-level visual features, including speed and direction of motion, and spatial extent, frequency, and orientation. New visual stimuli, relative to repeated visual stimuli, evoke activity in early visual regions, in both adults and infants (45–47). Thus, under the hypothesis that differences in stimulus features like visual orientation, motion, and frequency underlie the VOE response, we might expect to observe greater activity to unexpected than expected events, in both domains, in early visual regions, like V1 and MT.

If endogenous curiosity underlies the VOE response, a distinct set of regions would be recruited. Regions within the multiple demand (MD) network (48), including regions in the frontal and parietal cortices, the insula, and the anterior cingulate cortex, respond with greater activity when human adults are engaged in a range of difficult vs easy tasks, regardless of the task's modality (e.g. auditory vs visual) or content (e.g. verbal arithmetic vs motor inhibition). These regions are also engaged when people consider curiosity-inducing trivia questions (49), watch magic tricks (43), and learn from prediction error over rewards (50). Studies of infants show similar effects: Regions along the lateral surface of the frontal and prefrontal cortices show greater activity to violations of a previously learned visual or auditory pattern (51–53). Thus, if domain-general endogenous attention underlies the VOE response, then we would expect regions in the multiple demand network to respond with greater amplitude to unexpected than expected events from both domains.

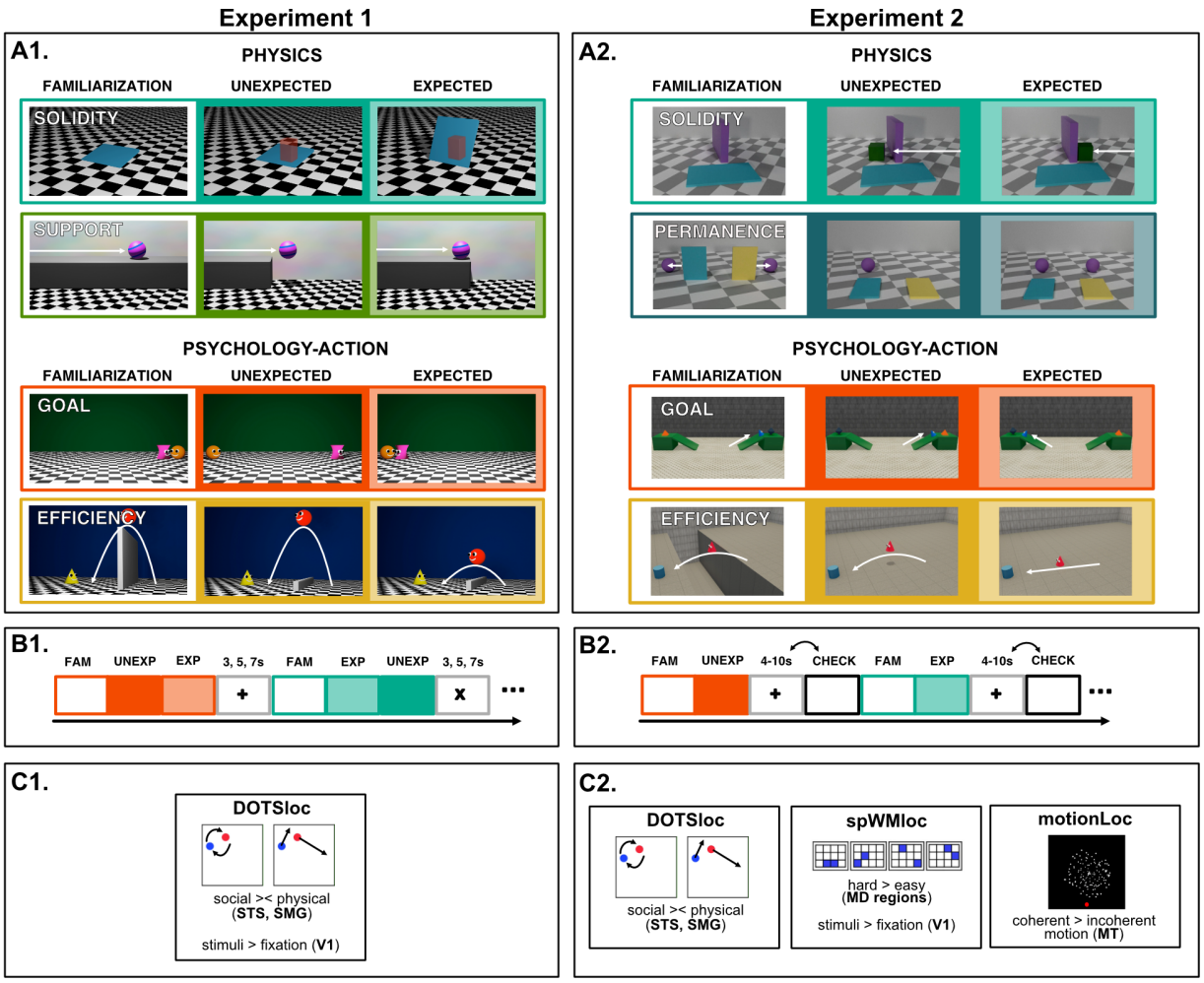


Figure 1. Overview of the methods of Experiments 1-2. (A1-2) Overview of VOE task. (top half) Stimuli from the domain of physics, including violations of object *solidity* and *permanence*. (bottom half) Stimuli from the domain of psychology, where the source of the violation is the agent performing a surprising action (psychology-action, including violations of *goal*-directed action and action *efficiency*). See Figure 5 for stimuli involving surprising physical outcomes, resulting from an agent's action, that were also included in Experiment 2. (B1-2) Structure of VOE run, with each trial containing a familiarization movie followed by both an expected or unexpected movie (Experiment 1), or an expected or unexpected movie (Experiment 2). (C1-2) Localizer tasks and contrasts for physics and psychology regions (interacting dots localizer, DOTSloc), multiple demand regions (spatial working memory localizer, spWMloc), and early visual regions (motionLoc).

Overview of current research

Here, we sought complementary evidence to the debate about infant VOE effects, by scanning the brains of adults while they watched events that were designed to test for physical and psychological expectations in infants. We studied cortical regions likely to be involved in the hypothesized processes underlying the VOE response (psychological and physical prediction, early visual processing, and goal-directed attention; see Figure 2A) in subject-specific functional regions of interest (ssfROIs), defined using validated localizer tasks from prior literature (35, 48, 54). See Methods section for details about our localizer tasks, and the ssfROI approach. We then measured the responses of these regions to unexpected and expected

psychological and physical events designed for infant studies. We tested whether the responses in each region are driven by manipulations of domain (psychology versus physics), event type (expected versus unexpected), or an interaction of these factors. Under domain-specific hypotheses, we expect a specific interaction between domain and event, with putative physics regions responding more to unexpected than expected physical events, but not psychological events, and vice versa for putative psychological regions. Under domain-general hypotheses, we expect greater responses to unexpected events for both domains in the same regions. We conducted two pre-registered functional magnetic resonance imaging (fMRI) experiments (see Methods for links to registrations). Here, we report the results of exploratory analyses from Experiment 1, which we pre-registered as confirmatory analyses in Experiment 2. Because the experiments and their results are similar, we report the methods and results folding across experiments. Conducting two experiments allows us to evaluate the robustness of our findings; thus, we will make the strongest claims about findings that replicate in both samples, generalizing across stimulus materials and design choices.

Our approach has both strengths and weaknesses. Studying adult brains, rather than infant brains, allows us to identify regions involved in each hypothesized process in individual participants using independently validated localizer tasks. This procedure gives us more confidence that the responses we measure correspond to the hypothesized mental processes, strengthening our "reverse inference" from neural activity to cognitive mechanisms (55, 56). Since there is a strong correspondence between the large-scale topography of adult brain networks between adults and infants, as early as they can be measured (57, 58), insights from adult brains could directly guide future studies of infant brains. However, researchers remain divided on how much continuity there is between infant and adult brains (59). We will return to the strengths and weaknesses of our approach in the discussion.

Results

We scanned the brains of 49 adults (N=17 in Experiment 1, N=32 in Experiment 2; see Methods for details) using fMRI while participants watched movies adapted from infant behavioral research, as well as one (Experiment 1) or three (Experiment 2) previously validated localizer tasks designed to identify regions involved in domain-specific psychological and physical prediction, low-level visual processing, and domain-general endogenous attention. See Methods for details about our localizer tasks.

Our violation-of-expectation (VOE) stimuli from Experiment 1 consisted of 4 handcrafted sets ('scenarios') of animated videos, adapted directly from previous studies from the infant cognition literature, involving violations of goal-directed action (*goal*) (5), action efficiency (*efficiency*) (60), object solidity (*solidity*) (61) and object support (*support*) (3).

Our violation-of-expectation (VOE) stimuli from Experiment 2 were selected from 2 large-scale procedurally generated video datasets, inspired by the infant cognition literature (1, 2), and also contained 3 hand-animated scenarios from Experiment 1. In total, there were 28[1] scenarios. The 12 scenarios from the domain of physics featured inanimate objects, barriers, and rotating fans. In surprising events, solid objects passed through each other (*solidity*) or blipped in and out of existence (*permanence*) (62) (Figure 1A). The 16 scenarios from the domain of psychology featured agents moving in physical environments, around physical obstacles, towards goal objects (Figure 1B-C), and were further divided into scenarios involving surprising *actions* (12 scenarios; Figure 1B), or surprising *environments* (4 scenarios; Figure 1C) in which the actions

[1] This deviates from our pre-registration which specified 32 scenarios, due to an error in our experimental scripts.

occurred. In the psychological scenarios involving surprising actions, agents changed their goals (*goal*), or acted inefficiently (*efficiency*) (Figure 1B). In the psychological scenarios involving surprising environments, agents moved through an (apparently) solid wall (*agent-solidity*) (16), or moved as though they were circumventing an obstacle, which was then missing (*infer-constraint*) (63). Our primary analyses focus on the psychology-action events; in further exploratory analyses, we studied neural responses to the psychology-environment events. Expected and unexpected events within each domain were matched along an array of low-level visual features (Figure S4). Independent adult observers rated the unexpected events from these three categories (physics, psychology-action, and psychology-environment) as equally surprising (Figure S5).

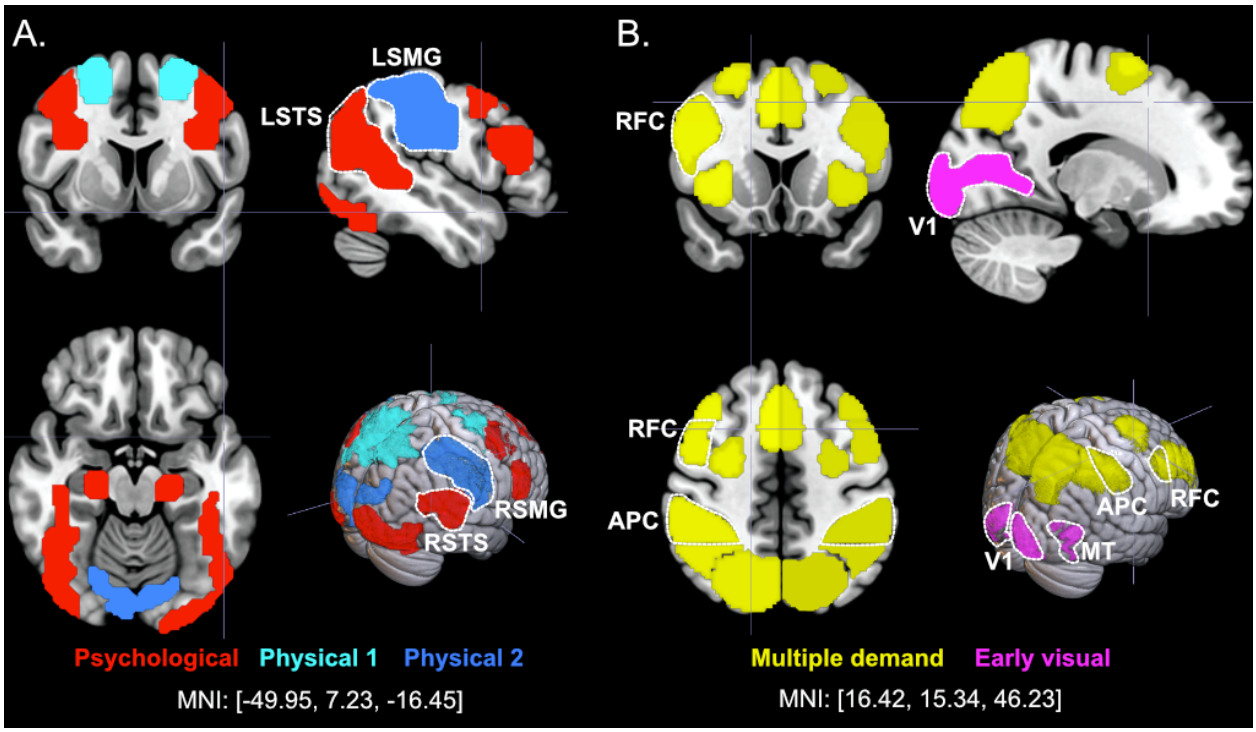


Figure 2. (A) Domain-specific and (B) domain-general parcels studied in Experiments 1-2, overlaid on an MNI152 template brain. Dotted lines indicate focal regions, pre-registered in Experiment 2, including left and right supramarginal gyrus (LSMG, RSMG), left and right superior temporal sulcus (LSTS, RSTS), right frontal cortex (RFC), anterior parietal cortex (APC), primary visual cortex (V1), and middle temporal area (MT). (A) The full set of domain-specific regions we explored, including frontoparietal parietal regions implicated in physical understanding, and frontal regions implicated in action observation. (B) The full set of domain-general regions we explored, including more multiple demand regions. MNI coordinates identifying the X, Y, and Z slice positions are listed below each figure. All data used to identify these parcels were independent of the data used to extract responses in the primary VOE task in both experiments (see SI for details about parcel definition).

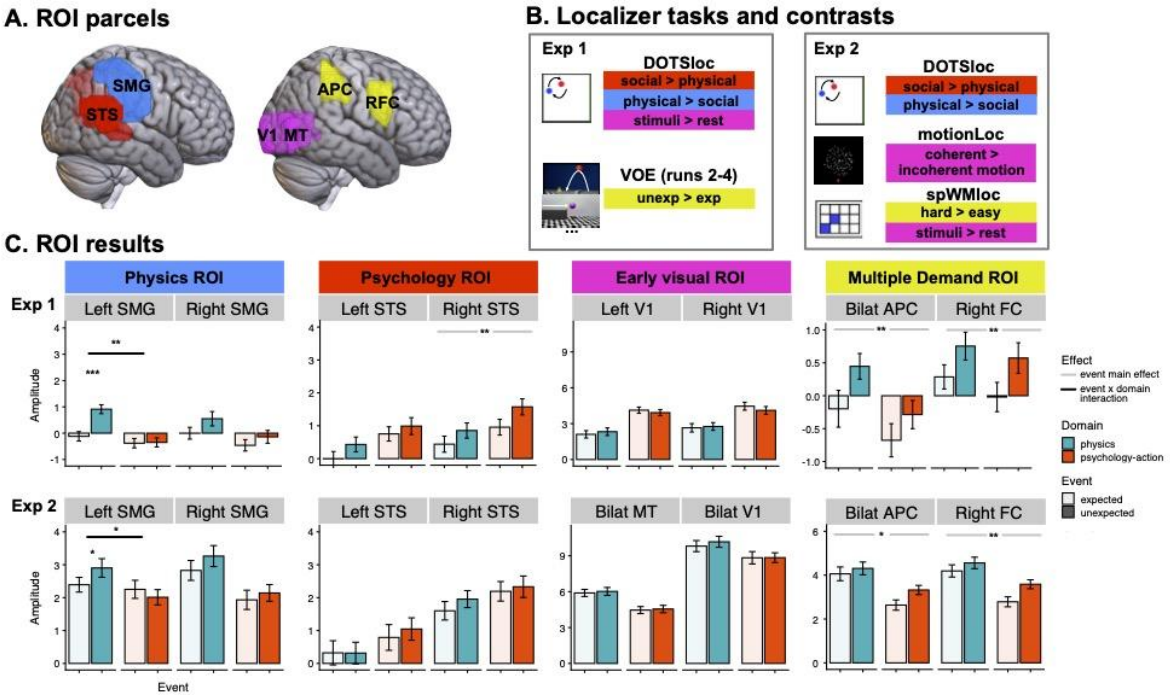


Figure 3. Results of univariate subject-specific functional regions of interest (ssfROI) analysis from Experiment 1 (exploratory) and Experiment 2 (confirmatory). (A) Parcels for all focal regions of interest (ROIs). (B) Localizer tasks and contrasts for voxel selection for both experiments. (C) ssfROI results in domain-specific regions (first two columns: left and right superior temporal sulcus, STS, and left and right supramarginal gyrus, SMG), domain-general early visual regions (bilateral primary visual cortices, V1, and bilateral motion-sensitive area, MT), and domain-general multiple demand regions (bilateral anterior parietal cortices, APC, and right frontal cortex, RFC). Y axis indicates the average beta (i.e. amplitude of response) per region, relative to fixation/rest, across 17 participants (Exp 1) and 32 participants (Exp 2). Error bars indicate the standard error of the mean, taking into account within-subjects variance.

Results in focal domain-specific regions

In a first set of analyses, we studied neural responses in a small number of regions that served as proxies for each of our hypothesized cognitive processes (psychological and physical prediction, early visual processing, endogenous attention). Our domain-specific physics regions were left and right supramarginal gyrus (SMG). Our domain-specific psychology regions were left and right superior temporal sulcus (STS). Both regions were chosen based both on their domain-specific functions based on prior literature, and on their preferential responses to social and physical stimuli from both our localizer task and our primary VOE task. Data used to choose these regions, and to select ROIs for individual participants, were independent of the data used to evaluate their responses to the VOE stimuli. All focal ROIs, including SMG and STS, were pre-registered ahead of Experiment 2. See SI for details.

We first conducted (Exp 1) and pre-registered (Exp 2) a manipulation check to assess whether the neural VOE effect declined across experimental runs (see SI for details). Following this procedure, we restricted our analyses to the first run of the VOE task in Experiment 1, and the first two runs of the VOE task in Experiment 2.

Then, we tested the hypothesis that the VOE response is supported in part by domain-specific processing. Throughout our results we will refer to neural VOE effects (i.e. unexpected vs expected) as “event” effects, and neural domain effects (i.e. psychology vs physics) as “domain effects”. Do we find evidence for domain-specific processing of violations of physical and psychological expectations, in cortical regions selective for those domains?

Physics ROIs

In Experiment 1, we first confirmed the selectivity of left and right SMG for physical over social stimuli: Both left and right SMG responded preferentially to physical events (left SMG: 95% CI = [0.252, 0.432], unstandardized B coefficient = 0.342, p-value < .001, two-tailed, Cohen’s d = 0.454, Bayes Factor (BF) > 1000; right SMG: [0.169, 0.365], B = 0.267, p < .001, two-tailed, d = 0.326, BF > 1000). Then, we conducted the key test for physical prediction error. We found that left SMG showed a VOE response that differed across domains (domain x event interaction effect: [0.104, 0.397], B = 0.25, p = 0.001, two-tailed, d = 0.422, BF = 2.758). LSMG responded more to unexpected than expected physical events (B = 1.031, p = <.001, two-tailed), but did not distinguish between unexpected and expected psychological events (B = 0.03, p = 0.888, two-tailed). RSMG did not show a main effect of event ([0.039, 0.399], B = 0.219, p = 0.018, two-tailed, d = 0.3, BF = 0.231), nor an interaction between event and domain ([-0.123, 0.237], B = 0.057, p = 0.535, two-tailed, d = 0.078, BF = 0.017).

We then pre-registered the prediction for domain-specific prediction error in left SMG in Experiment 2. We again found that left SMG showed a signature of domain-specific prediction error: an interaction between domain and event ([0.104, 0.412], B = 0.258, p = 0.001, two-tailed, d = 0.441, BF = 2.185), with greater responses for unexpected than expected physical events B = 0.51, p = 0.023, two-tailed), and no significant VOE effect for psychological events (B = -0.241, p = 0.28, two-tailed). Right SMG showed a marginally higher response to unexpected events regardless of domain ([-0.007, 0.329], B = 0.161, p = 0.062, two-tailed, d = 0.253, BF = 0.078), with no interaction between event and domain ([-0.111, 0.224], B = 0.057, p = 0.511, two-tailed, d = 0.089, BF = 0.017). Like in Experiment 1, both left and right SMG responded more to physical than psychological events (left SMG: [0.104, 0.412], B = 0.258, p = 0.001, two-tailed, d = 0.441, BF = 2.185; right SMG: [0.336, 0.672], B = 0.504, p < .001, two-tailed, d = 0.79, BF > 1000). See Methods and SI for details about model specification.

Psychology ROIs

In Experiment 1, we found that both left and right STS responded more to psychological than physical events (left STS: [-0.509, -0.148], B = -0.329, p < .001, two-tailed, d = -0.449, BF = 6.951; right STS: [-0.699, -0.317], B = -0.508, p < .001, two-tailed, d = -0.654, BF > 1000). However, we did not find evidence for a distinctively psychological prediction error–an interaction between event and domain– in these regions (left STS: [-0.131, 0.23], B = 0.05, p = 0.593, two-tailed, d = 0.068, BF = 0.016; right STS: [-0.218, 0.116], B = -0.051, p = 0.553, two-tailed, d = -0.075, BF = 0.016) Instead, we found that the right STS responded more to unexpected events from both domains (right STS: [0.091, 0.425], B = 0.258, p = 0.003, two-tailed, d = 0.381, BF = 1.177 whereas the left STS did not show a significant main effect of event ([-0.012, 0.349], B = 0.168, p = 0.07, two-tailed, d = 0.23, BF = 0.074)

In planning for Experiment 2, in which we plausibly had greater statistical power (due to the larger sample size, more stimuli, and more runs of data), we pre-registered two alternative hypotheses: that the STS would show domain-specific psychological prediction error, which would lead to an interaction between event and domain, or that the STS encodes both physical and psychological information relevant for action understanding, which would lead to a main effect of domain, and of event, but no interaction effect. In the confirmatory analyses of

Experiment 2, we found support for neither hypothesis. Both left and right STS responded more to psychological events (left STS: [-0.491,-0.109], B=-0.3, p=0.002, two-tailed, d = -0.413, BF = 1.584; right STS: [-0.405,-0.08], B=-0.242, p=0.004, two-tailed, d = -0.392, BF = 0.861). However, neither left nor right STS responded more to unexpected than expected events (left STS: [-0.129,0.253], B=0.062, p=0.524, two-tailed, d = 0.086, BF = 0.019; right STS: [-0.039,0.286], B=0.123, p=0.139, two-tailed, d = 0.2, BF = 0.039), and there was no interaction between domain and event in these regions (left STS: [-0.257,0.125], B=-0.066, p=0.501, two-tailed, d = -0.091, BF = 0.019; right STS: [-0.109,0.216], B=0.054, p=0.517, two-tailed, d = 0.087, BF = 0.016). Results were similar when we defined STS ROIs not based on the external localizer, but rather, as voxels that responded more to psychological than physical VOE events (see SI for details). Thus, we did not find consistent evidence for domain-general or domain-specific psychological prediction error in our focal psychology ROIs.

Next, we tested for evidence for domain-general processing of violations of expectation, in cortical regions associated with visual processing and endogenous attention.

Early visual ROIs

In Experiment 1, we found via exploratory analyses that neither left nor right V1 responded more to unexpected than expected events (left V1: [-0.222, 0.23], B = 0.004, p = 0.973, two-tailed, d = 0.004, BF = 0.018; right V1: [-0.293, 0.174], B = -0.06, p = 0.618, two-tailed, d = -0.063, BF = 0.021). Both left and right V1 responded more to psychological events (left V1: [-1.13, -0.678], B = -0.904, p < .001, two-tailed, d = -0.986, BF > 1000; right V1 [-1.022, -0.555], B = -0.788, p < .001, two-tailed, d = -0.832, BF > 1000).

In Experiment 2, we found again that neither bilateral V1 nor bilateral MT responded differently to unexpected and expected events (V1: [-0.171, 0.356], B = 0.093, p = 0.492, two-tailed, d = 0.093, BF = 0.027; MT: [-0.079, 0.187], B = 0.054, p = 0.428, two-tailed, d = 0.107, BF = 0.015). Both bilateral V1 and bilateral MT responded more to physical than psychological events (the opposite effect from that in Experiment 1) (V1: [0.312, 0.839], B = 0.575, p < .001, two-tailed, d = 0.575, BF = 145.691; MT: [0.594, 0.86], B = 0.727, p < .001, two-tailed, d = 1.437, BF > 1000). The higher average response to physical events in MT appears to be driven by variance in low-level statistics in the stimuli (see SI for details); after controlling for these features, MT no longer showed a significant domain preference ([-0.01, 0.389], B =0.19, p = 0.064, two-tailed, d = 0.137). V1 continued to show a preference for physical events, after accounting for these same features ([0.122, 0.813], B =0.468, p = 0.008, two-tailed, d = 0.196)

In sum, we found no consistent domain-specific responses, and consistently found no VOE effects, in early visual regions.

Goal-directed attention ROIs

Lastly, we tested the hypothesis that the VOE response is (also) supported by domain-general endogenous attention by studying responses in two multiple demand regions: the right frontal cortex (RFC) and bilateral anterior parietal cortex (APC; see SI for evidence for low overlap with SMG ROIs in individual participants). These two particular ROIs were pre-registered ahead of Experiment 2, and chosen based on prior literature and the results of Experiment 1 (see SI for details).

In Experiment 1, we found via exploratory analyses that right frontal cortex (RFC) responded more to unexpected than expected events (main effect of event: [0.097,0.434], B=0.265, p=0.002, two-tailed, d = 0.387, BF = 1.381). This region did not respond preferentially to physical or psychological events (main effect of domain: [-0.047,0.291], B=0.122, p=0.16,

two-tailed, d = 0.178, BF = 0.036), and there was no interaction between event and domain ([-0.2,0.138], B=-0.031, p=0.719, two-tailed, d = -0.045, BF = 0.014). We found that bilateral anterior parietal cortex (APC) also responded more to unexpected than expected events (main effect of event: [0.096,0.422], B=0.259, p=0.002, two-tailed, d = 0.391, BF = 1.453), and responded more to physical than psychological events (main effect of domain: [0.14,0.466], B=0.303, p<.001, two-tailed, d = 0.458, BF = 7.935), with no interaction between domain and event ([-0.099,0.227], B=0.064, p=0.447, two-tailed, d = 0.096, BF = 0.017).

We then pre-registered these same predictions in Experiment 2. We found that both RFC and APC responded more to unexpected than expected events (RFC: [0.103, 0.48], B = 0.291, p = 0.003, two-tailed, d = 0.407, BF = 1.36; APC: [0.032, 0.436], B = 0.234, p = 0.024, two-tailed, d = 0.305, BF = 0.208), with no intersection between domain and event (RFC: [-0.297, 0.08], B = -0.109, p = 0.261, two-tailed, d = -0.152, BF = 0.028; APC: [-0.312, 0.092], B = -0.11, p = 0.287, two-tailed, d = -0.144, BF = 0.029). Both RFC and APC responded more to physical than psychological events (RFC: [0.407, 0.783], B = 0.595, p < .001, two-tailed, d = 0.831, BF > 1000; APC: [0.401, 0.804], B = 0.602, p < .001, two-tailed, d = 0.786, BF > 1000).

Controlling for visual features

We tested in an exploratory analysis whether any results (domain-specific event response in LSMG, domain-general event responses in APC and RFC) from Experiment 2 are explained by variability in the lower-level visual statistics in our stimuli (e.g. motion, spatial extent). We focused this analysis on Experiment 2 which had many more unique stimuli than Experiment 1 and therefore could support the full set of visual features as predictors. We found that after accounting for variability in the contrast, luminance, motion, spatial frequency content, rectilinearity, and curvilinearity of the stimuli, all positive and negative VOE effects from our confirmatory (Exp 2) analyses held, including the null findings in V1, MT, LSTS, and RSTS, as well as the positive effects in LSMG, APC, and RFC. The domain preferences in three domain-specific regions (RSMG, LSTS, RSTS) also remained significant, after controlling for the visual features. The domain preferences for V1 held after controlling for these features, but the direction of these preferences were inconsistent across experiments and stimuli. The apparent preference for physical events in the two MD regions (APC and RFC) and one visual region (MT) were no longer significant, after controlling for visual features. See SI for details.

Exploring domain and event effects across many cortical regions

In the univariate analyses over focal regions, we searched for neural VOE effects in 8 domain-specific and domain-general regions. However, we also wanted to characterize the responses of regions across the cortex. As a complementary approach, in further exploratory analyses, we studied domain and event univariate effects in a larger set of 18 domain-specific regions and 24 domain-general regions (see SI for details). We studied the responses in these regions in two ways. First, we looked in each region for evidence of a domain or event effect in the univariate response, with a conservative significance threshold to account for the number of regions we explored (Bonferroni correction; $\alpha = .002$ for domain-general regions; $\alpha = .003$ for domain-specific regions). None of the regions we explored, in either experiment, showed a significant VOE effect, though many showed differential responses to physical and psychological events. See SI for details about this analysis, as well as results from whole-brain analyses.

Finally, we conducted a series of analyses investigating the reliability of event and domain information across domain-specific and domain-general regions. We took advantage of the 2x2 design of our experiment (psychology vs physics; unexpected vs expected), split the data into two halves, and computed 2 effect sizes per split: (i) domain preferences for expected events,

and separately, domain preferences for unexpected events, and (ii) event preferences for psychological events, and separately, event preferences for physical events. Then, across regions, we studied the reliability of the effect sizes for events across domains, and domains across events. Are domain-general regions and domain-specific regions organized by domain and event, respectively? Or do the responses in these regions go beyond the information they were defined over (attentional demand and visual processing for domain-general regions; social vs physical prediction for domain-specific regions)?

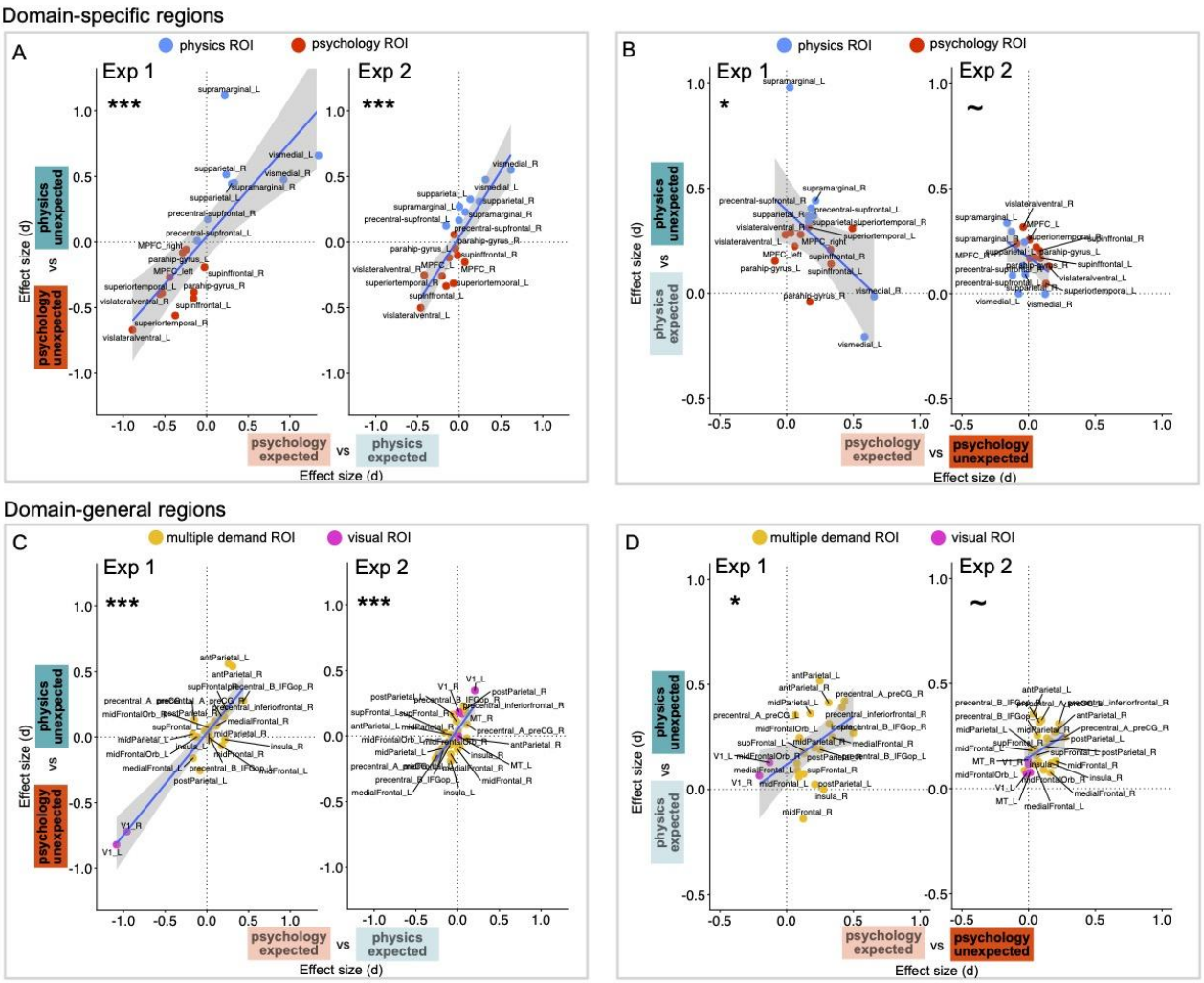


Figure 4. Univariate effect size results across all domain-specific regions (A-B) and domain-general regions (C-D) from Experiments 1-2. (A) and (C) show correspondence between domain information across event types. (B) and (D) show correspondence between event information across domains. Effect sizes from Experiment 2 were derived from models that controlled for low-level visual statistics between events. ~ $p < .10$, * $p < .05$, ** $p < .01$, ***$p < .001$, two-tailed, non-parametric test for independence.

Across both Experiments 1 and 2, we found that response magnitude across 24 putatively domain-specific regions is reliable by domain, but not by event. For these regions, the size of a region's domain effect (psychology vs physics) for expected events strongly predicts the size of the same region's domain effect for unexpected events (Exp 1: $r = 0.782$, $p < .001$; Exp 2: r =

0.844, p < .001). However, the size of a region's VOE effect (unexpected vs expected) for psychology-action events was weakly anticorrelated with the size of the same region's VOE effect for physics events, for both experiments (Exp 1: r = -0.495, p = 0.04; Exp 2: r = -0.419, p = 0.087). The reliability of domain information was greater than for event information (bootstrapped difference in correlations; Exp 1: 95% CI [0.722, 1.553], p < .001; Exp 2: 95% CI [0.894, 1.688], p < .001). See Figure 4A-B.

What about domain-general regions, that were defined based on responses to visual information (V1 and MT) or to a spatial working memory task (MD regions), with no reference to domain information? We found that these regions' responses were reliable for both domain and event contrasts, in both Experiments 1 and 2. Across these regions, the domain effect (psychology vs physics) for expected events strongly predicted the domain effect for unexpected events (Exp 1: r = 0.892, p < .001; Exp 2: r = 0.739, p < .001). In addition, the psychology event effect (unexpected vs expected) positively predicted the physical event effect (Exp 1: r = 0.478, p = 0.021; Exp 2: r = 0.377, p = 0.063). Like in domain-specific regions, the reliability of domain information was greater than the reliability of event information (bootstrapped difference in correlations; Exp 1: (95% CI [0.084, 0.558], p =.004); Exp 2: 95% CI [0.268, 0.989], p < .001). See Figure 4C-D.

Multivariate tests of event and domain information

In addition to these univariate analyses, we pre-registered and conducted a series of multivariate pattern analyses (MVPA). We tested whether any of our focal regions contained distinct patterns of activity for unexpected vs expected events (and if so, whether these patterns were domain-specific or domain-general). By contrast to the univariate results, we found no evidence for a consistent spatial pattern distinguishing unexpected vs expected events in any region, and at the same time, strong evidence for consistent spatial patterns distinguishing between domains in many of our focal regions. This null result held even though we used Euclidean distance as the distance metric, which incorporates differences in response magnitude. The dissociation between univariate and multivariate information was unexpected to us; we will speculate about possible implications of this result in the General Discussion. The full multivariate results are presented in the SI.

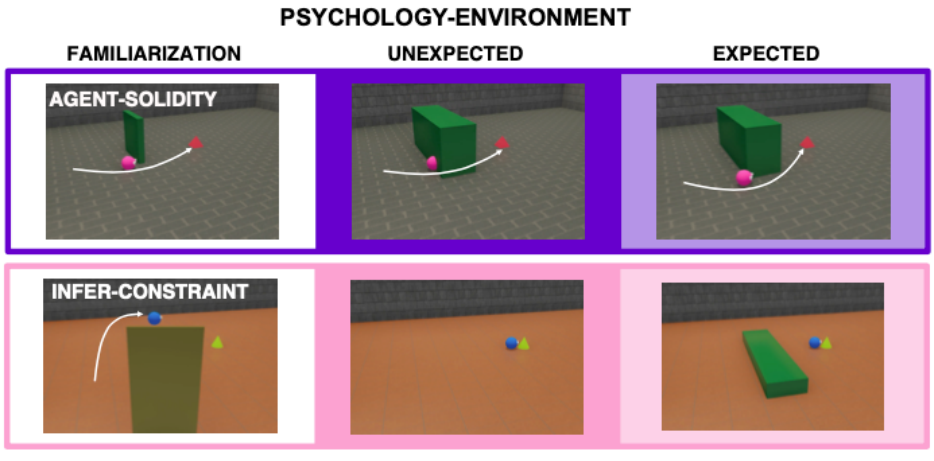


Figure 5. Stimuli from the domain of intuitive psychology, wherein the actions of an agent revealed a surprising physical outcome in the surrounding environment (psychology-environment). In *agent-solidity*, an agent passes through a solid wall; in *infer-constraint*, an obstacle that explains an agent's action is missing.

VOE towards physically surprising outcomes, revealed by an agent

In our primary analyses, reported above, we studied neural responses to surprising actions (psychology-action events). How do our focal domain-specific and domain-general regions respond to surprising events involving both agents and objects (psychology-environment scenarios, Figure 1C), wherein a physical outcome is rendered surprising in light of an agent's action? To ask this question, in exploratory analyses we modeled the responses of all the focal regions in the 4 psychology-environment expected and unexpected scenarios. When restricting the analysis to data from the first 2 runs, like in our confirmatory analyses in psychology-action and physics events, the only focal ROI that showed a VOE effect was the right STS ([0.095, 0.542], B = 0.319, p = 0.006, two-tailed, d = 0.576, BF = 1.156).

Further exploratory analyses over all runs of the experiment suggested that many focal ROIs showed strong VOE effects towards these stimuli across runs, including domain-specific physics ROIs (left SMG: [0.229, 0.469], B = 0.349, p < .001, two-tailed, d = 0.492, BF > 1000; right SMG: [0.156, 0.415], B = 0.285, p < .001, two-tailed, d = 0.372, BF = 65.936), a domain-specific psychology ROI (right STS: [0.159, 0.342], B = 0.25, p < .001, two-tailed, d = 0.462, BF > 1000) and MD ROIs (APC: [0.034, 0.293], B = 0.163, p = 0.014, two-tailed, d = 0.213, BF = 0.146; RFC: [0.079, 0.34], B = 0.209, p = 0.002, two-tailed, d = 0.271, BF = 0.934), though not early visual regions (bilateral V1: [-0.064, 0.203], B = 0.07, p = 0.306, two-tailed, d = 0.088, BF = 0.012; bilateral MT: [-0.009, 0.175], B = 0.083, p = 0.078, two-tailed, d = 0.152, BF = 0.023), nor left STS ([-0.084, 0.209], B = 0.063, p = 0.404, two-tailed, d = 0.072, BF = 0.011). These effects were spatially consistent across participants, appearing in the SMG and STS in whole-brain random effects analyses. See Figure S11.

General Discussion

Why do infants look when a ball (apparently) passes through a solid wall? The underlying mental processes that guide looking to events like these remain controversial, despite decades of behavioral studies. Thus in the current work, we used the tools of cognitive neuroscience to directly and simultaneously examine these mental processes, albeit in adults. We localized the brain regions in individual adult participants that support domain-specific and domain-general processes hypothesized to account for VOE (domain-specific psychological and physical reasoning, domain-general visual prediction error, and domain-general task-driven attention), and tested which of these processes show a corresponding neural VOE effect for stimuli from classic infant experiments. Overall, we found evidence that unexpected events in these stimuli (i) did not evoke processes similar to early-stage visual processing, (ii) evoked processes similar to endogenous goal-driven attention, and (iii) for physical events, evoked domain-specific distinctively physical processing, in adult brains.

Before we discuss our positive findings, let us consider the implications of our negative findings from early visual regions. We found no evidence for the hypothesis that VOE stimuli evoke responses associated with visual processing of novel visual features. Primary visual cortex (V1) and motion-sensitive area (MT), did not respond more to unexpected than expected VOE events: The voxels that, in individual participants, were maximally responsive to visual stimuli (in V1) or to coherent motion (in MT), responded equally to unexpected and expected scenarios, regardless of domain. This result provides some evidence against the hypothesis that unexpected events in infant studies attract attention merely because they contain an array of novel low-level visual features (11, 12), because V1 and MT should be sensitive to exactly these features.

Next, we consider the implications of the findings from domain-specific physical and psychological regions, and domain-general multiple demand regions, for VOE in adults.

Violations of physical expectations

What happens in the minds and brains of adults when they see a violation of object support, solidity, or permanence? Prior research proposes that adults possess a system for 'intuitive physics' (64): a capacity to represent the visual world in terms of the objects and surfaces in it, including inductive biases that objects are permanent and solid, that allows adults to form expectations about what will happen next, and to detect deviations from those expectations. Prior work suggests both a distinctive neural source of these capacities (33–35), as well as early emergence in infant behavior (10). The current results suggest that when adults see a physically surprising event, it evokes both a prediction error within that system for intuitive physics (supported by the frontoparietal physics network, including the SMG), and also a domain-general orienting response towards that event (supported by multiple demand regions, including the APC and RFC).

Whereas in behavioral research, domain-general and domain-specific contributions to the VOE effect are difficult to separate, neuroimaging allowed us to identify both domain-specific and domain-general correlates of the VOE effect simultaneously. Having found evidence for both processes, many questions follow. One question is whether physical prediction error is initially calculated in one region, and passed to the other(s), and if so in which direction. Does a physical prediction error signal arise initially in LSMG, which is then read out by RFC or APC? Or does the LSMG pass physically relevant information to MD regions, and then receive a signal of physical prediction error from these regions? These questions are best addressed using neuroimaging techniques with good temporal resolution, like electroencephalography (EEG) and magnetoencephalography (MEG).

Another question is whether the neural population code for unexpectedness in MD regions, like RFC and APC, is truly domain-general. When we measured the reliability of domain and event univariate responses in domain-general regions, we found that across 24 regions, the strength of an MD region's VOE effect generalized (albeit weakly) across domains. However, we could not test whether the pattern of response to physically unexpected events could be used to decode psychologically surprising events, or vice versa, because we could not measure reliable spatial patterns distinguishing expected versus unexpected events. By contrast, there were consistent patterns of information distinguishing the physical and psychological events, both within and across event types, in many regions (see SI for details). In sum, unexpected events led to greater activity in MD regions, but not in a consistent spatially structured manner. In this way, our results are consistent with prior evidence that prediction error increases response magnitude but reduces population code precision (65, 66). If this interpretation is true, then MVPA cannot be used to test hypotheses about the representations underlying VOE responses, at least the way they are conceptualized in the current research.

What can we infer or predict about infant brains, given these findings from adults? Studying the brains of adults to evaluate hypotheses about neural function and behavior in infants has both strengths and limitations. One strength was that studying adults allowed us to be more confident about the functions of the regions we studied, by using validated localizer tasks that targeted each candidate mental process underlying VOE. This design was possible because adults can tolerate long scans and can be instructed to perform tasks in the scanner. It is much harder to design and run localizer tasks in infants, but without localizers, reverse inference over functional activation alone is not straightforward (55, 56) (e.g. in the APC and SMG, which occupy approximately the same cortical territory on average, but are spatially and functionally distinct in

individual adult participants; see SI). In our experimental design, we prioritized stimuli and procedures that exactly correspond to prior studies of infants. A weakness of this strategy was that these may have not been the ideal conditions for maximizing adult engagement: adults' neural VOE effect quickly habituated over just a few experimental runs (see SI for details).

Prior neuroimaging studies suggest that infants have similar organization of large-scale cortical networks, as well as similar cortical responses evoked by agents and objects, to those of adults (36, 39, 57, 58, 67–69). Thus we speculate that all the focal regions we studied in adults are present in approximately the same locations, and functional, in infants under one year of age. Most relevantly, work using near-infrared spectroscopy in 5- to 7-month-old infants reported increases in activity in parietal cortex when infants saw objects move in a discontinuous path, or change speed (38). If multiple demand and frontoparietal physics regions could be interrogated separately and studied in infant brains, then we predict that violations of physical expectations would evoke activity in domain-specific and domain-general regions in infants, just as in adults, and that both would contribute to infant looking behavior in VOE studies.

Violations of psychological expectations

How do the human adult mind and brain respond to deviations from efficient or goal-directed action? In addition to capacities for physical understanding, prior research shows that adults have an intuitive theory of rational action (70): a capacity to represent people as agents with mental states who plan intentional actions at a cost to themselves, which allows adults to predict and explain other agents' behaviors. These capacities emerge in infancy (6, 9), and are likely supported by cortical regions involved in action processing (71).

In the current study, apparently irrational actions evoked increased activity in regions localized by endogenous attention, suggesting that psychological prediction error, like physical prediction error, leads to a domain-general orienting response. However, the existence of domain-specific prediction error, and the role of the STS, were less clear. In our study, STS responded to the actions of agents, consistent with the social functions of the superior temporal sulcus (27–32). However, evidence of social prediction error in the STS was less conclusive. Prior literature is similarly mixed, with some researchers finding activation in the STS for violations of rational action (31, 32, 41, 72, 73), and others finding activation in frontoparietal regions that could reflect the same responses we measured in APC and RFC (40–42). Thus, while the STS is likely involved in the processing of social information more broadly, it is unclear how the STS is involved in expressing an intuitive theory of action, including prediction error over that theory, in particular. One possibility is that the STS does encode prediction errors over action, but shows a more sustained response for action outcomes that are harder to explain away (e.g from prior work, when a person opens a door with her knee, even though her hands are free (73); a person expressing disgust at an object, and then reaching for it (31)), whereas the actions we tested here and in prior work (42) were much simpler (someone changing their mind about which object to pursue; someone acting inefficiently) and easier to explain away, and thus led to a smaller STS VOE response.

As early as can be measured, activity in the superior temporal cortex is evoked by social stimuli in infants as well as adults, responding to faces (36, 74), actions (75), and social interactions (76–78). But, like in adults, infant STS may not encode action prediction errors during simple scenes involving violations of rational action. In one near-infrared spectroscopy (NIRS) study with 9-month-old infants, Southgate et al. (2014) (40) measured responses from the temporal and parietal cortex while infants watched an animated agent move towards one object, and then move towards the same object in a new location or move towards the same location, now occupied by a new object, much like our goals task. They found that two contiguous channels

over the left anterior parietal cortex responded to changes in an agent's goals; no other contiguous channels showed a similar response. Based on these observations, we predict that infant looking to VOE events involving surprising actions will reflect both domain-specific and domain-general neural sources, though it is an open question whether the STS in infants encodes these prediction errors.

Distinct and shared representations from intuitive physics and psychology

Our study found evidence supporting the broad division between the physical and psychological domains in the human brain. First, many of the 42 regions we studied preferentially responded to events involving agents or objects; this was true both for domain-specific regions we defined based on a social versus physical contrast, and also for domain-general multiple demand regions we defined based on spatial working memory (Figure 4A, 4C). Furthermore, we found that in domain-specific regions, VOE effect sizes tended to trade off between domains: Regions that tended to show a VOE effect in one domain tended not to show that effect in the opposite domain.

However, in some ways, our results also highlight the interactions between these two domains. First, physical outcomes that were surprising in light of observed actions evoked activity in both psychological and physical ROIs. These events plausibly required computations from both domains. It is not surprising, by itself, to see an agent move on a straight path, but it is surprising if that path is through a solid object. It is not surprising, by itself, to see an occluder reveal empty space, but it is surprising if an obstacle, implied by an agent's action, is not there. We suggest that computations from both domains are necessary for adults and infants to make sense of these events.

Because agents have physical bodies, act in a physical world, and their plans reflect information about that world, adults' and infants' understanding of even simple actions may require the integration of computations between physical and psychological domains. For example, representing the efficiency of an action may require first representing the agent and obstacle as solid bodies, and the agent as a body that can generate force against gravity. The best computational models of how infants understand other people's goal-directed actions contain a joint model for action planning and physical simulation (1). While we have followed a long tradition, from both cognitive neuroscience and developmental psychology, of studying intuitive psychology and physics as contrasting domains, our imposed labels may be obscuring common or linked representations that organize the functions of domain-specific regions like the STS and SMG. Future work could explicitly link the representations from computational models of early intuitive psychology and physics to neural responses to better understand our capacity to reason about agents acting in a physical world.

Methods

Open science practices

The methods and analyses of these experiments were pre-registered prior to data collection, including several updates. Our pre-registration documents, openly available at https://osf.io/sa7jy/registrations, detail all decisions and updates and the status of data collection and analysis. All experiment scripts, including stimuli shown to participants, as well as the data and analysis scripts required to reproduce statistical results, can be found at https://osf.io/sa7jy/. De-faced brain images from participants in Experiments 1 and 2 who consented to share them (N = 16 for Exp 1; N = 29 for Exp 2) will be shared on OpenNeuro prior to the publication of this paper.

Participants

We recruited 20 participants (Mean age = 25.1y, range = 19-45; 17 right-handed; 15 female, 5 male; 10 White; 10 Black, Asian, or Latine) for Experiment 1, and 33 participants (Mean age = 25.7y, range 18-45; 30 right-handed; 21 female, 12 male; 19 White; 14 Black, Asian, Latine, or multiracial) for Experiment 2, all from the Boston area. Two participants were excluded from Experiment 1 due to technical issues. One participant each was excluded from Experiment 1 and Experiment 2 for not contributing usable, low-motion fMRI data. This left a final sample of N=17 for Experiment 1, and N=32 for Experiment 2. Participants had normal or corrected-to-normal vision and no MRI contraindications. We chose the sample size for Experiment 2 using a combination of simulation power analyses over Experiment 1 (see pre-registration for details), and other considerations of time and cost. All study procedures were approved by the MIT Committee on the Use of Human Subjects. Participants were asked to provide written informed consent before participation, and were paid $30 per hour.

Data acquisition

For full scanner protocols for both experiments, please see our pre-registration documents at https://osf.io/sa7jy/. In brief, for both experiments, neuroimaging data were acquired from a 3-Tesla Siemens Magnetom Prisma scanner located at the Athinoula A. Martinos Imaging Center at the McGovern Institute at MIT, using the standard 32-channel head coil. Participants viewed stimuli projected to a 12” x 16” screen behind the scanner, at a visual angle of approximately 14 x 19 degrees, through a mirror. Participants underwent an anatomical scout scan (auto-align, acquired in 128 sagittal slices with 1.6mm isotropic voxels, used to identify key anatomical landmarks and position the bounding box for subsequent anatomical and functional scans; TA=0.14; TR=3.15ms; FOV=260mm), and a high-resolution MPRAGE anatomical scan (T1-weighted structural images acquired in 176 interleaved sagittal slices with 1.0mm isotropic voxels, TA=5:53, TR=2530.0ms; FOV=256mm, GRAPPA parallel imaging, acceleration factor of 2).

In Experiment 1, participants underwent 6 runs of functional scans (gradient-echo EPI sequence sensitive to Blood Oxygenation Level Dependent (BOLD) contrast in 3mm isotropic voxels in 46 interleaved near-axial slices covering the whole brain; EPI factor=70, TR=2s, TE=30.0ms, flip angle=90 degrees, FOV=210mm). Two of these runs were dedicated to the DOTS localizer task. The remaining 4 runs were dedicated to our primary VOE task of interest. In total, the scanning session lasted about 60 minutes.

In Experiment 2, participants underwent 10 runs of functional scans (gradient-echo EPI sequence sensitive to Blood Oxygenation Level Dependent (BOLD) contrast in 3mm isotropic voxels in 50 interleaved near-axial slices covering the whole brain; EPI factor=70; TR=2s; TE=30.0ms; flip angle=90 degrees; FOV=210mm). Six of these runs were dedicated to our 3

localizer tasks, two runs apiece. The remaining 4 runs were dedicated to our primary VOE task of interest. In total, the scanning session lasted about 90 minutes.

Overview of preprocessing and analysis of brain images

A detailed description of our neuroimaging analysis pipeline can be found in the SI. In brief, data were preprocessed using fMRIPrep (79), which included brain extraction, tissue segmentation, co-registration to MNI space, motion correction, and confound estimation. The preprocessed BOLD images were analyzed using custom lab scripts using custom scripts, which included run-level exclusion based on motion, first and second level modeling, and whole brain analysis. During run-level modeling, all regressors other than head movement parameters were convolved with a standard double-gamma hemodynamic response function, with a high pass filter applied to both the data and the model. Event regressors were defined as a boxcar from the start and end of each block (localizer tasks) or event (VOE task). These first-level general linear models (GLMs) were then passed to subject-level and group-level analyses.

Localizer tasks

Social versus physical interaction (DOTSloc)

The DOTSloc task (35) reliably evokes responses in the superior temporal sulcus (STS) and supramarginal gyrus (SMG) (ROIs for psychological and physical prediction). Stimuli consisted of 32 unique 10-s movies of two dots moving as though they are physical objects, or as though they are interacting socially. Participants watched the dots, imagined the trajectory of one of the dots when it disappeared briefly, and indicated whether the final position of the hidden dot matches what they imagined using a button press. Each run included 19 blocks (8 physical blocks, 8 social blocks, and 3 rest blocks). On social and physical blocks, participants saw two different videos from the corresponding condition. Participants saw two runs, except for two participants in Exp 2 who only underwent one run due to time restrictions. Each run lasted approximately 8.2 minutes. This task was also used to define the V1 ROIs in Experiment 1.

Spatial working memory (spWMloc)

The spWMloc task (48), openly available at https://evlab.mit.edu/funcloc/, identifies regions in the multiple demand (MD) network, including bilateral anterior parietal cortex, and right frontal cortex (ROIs for goal-driven attention). Stimuli were rectangular 8-by-8 grids. Participants saw a sequence of grid-squares change color, either one (easy condition) or two (hard condition) at a time. They were asked to remember the locations of the changing squares over the sequence, and indicated using a button press which of two alternative grids matched the resulting layout, with feedback. Participants saw two runs, except for one participant in Exp 2 who only underwent one run due to time restrictions. Each run included 20 16-second blocks (6 easy, 6 hard, and 4 rest blocks), and lasted approximately 7.5 minutes. This task was also used to define the V1 ROI in Experiment 2.

Motion (motionLoc)

The motionLoc task (54) identifies motion-sensitive regions (MT) (ROI for early visual processing). This task contrasts coherent vs random dot motion to enable the identification of motion-sensitive voxels. Participants fixated on a red dot near the bottom center of the screen while a large circular space of small moving dots played above fixation. The dots either moved coherently (in a uniform direction) or randomly around the space. Participants pressed a button whenever the red dot flickered. Participants saw two runs, except for two participants in Exp 2 who only underwent one run due to time restrictions. Each run lasted approximately 4.6 minutes.

Primary VOE task: Experiment 1

Each VOE run had an event-related design: 8 trials (2 apiece of the solidity, support, goal, and efficiency scenarios), with jittered fixation/attention check periods of 3, 5, or 7 seconds in between each trial, and then a final rest period, lasting a total of approximately 8.0 minutes. All participants saw 4 runs. Each trial had 3 parts: a familiarization movie followed by two test movies presented in random order (expected and unexpected). All movies lasted 6s with a 250ms interstimulus interval, and each movie played twice in a row each time it was presented, followed by a jittered fixation/attention check. Participants were asked simply to pay attention to the movies. During the fixation period, participants pressed a button if the fixation cross was the letter X instead of a plus symbol (+) (33% of trials). The stimuli flipped horizontally for half of the trials to introduce minor visual variability across the run.

Primary VOE task: Experiment 2

Each VOE run had an event-related design: a 10s rest period, 16 trials (6 physics, 6 psychology-action, and 4 psychology-environment), and then a final 10s rest period, lasting a total period of approximately 7.0 minutes. All participants saw four runs, except for one participant in Exp 2 who only underwent three runs due to time restrictions. Each trial had 4 parts: a familiarization movie (7.5s), a corresponding test movie (7.5s; either unexpected or expected), each followed by a 250ms interstimulus interval, a fixation cross for a jittered duration of 4-10s, and an attention check (2s). Participants were asked to pay attention to the movies. During the attention check, they saw an image of an agent, object, or surface texture, and responded via button press to indicate whether that image appeared in the most recent trial. In anticipation that we may need to restrict our analysis to the first 2 runs, scenarios were split into two halves, one half assigned to runs 1-2 and the other assigned to runs 3-4, so that analyses over the first two runs would be conducted on the same stimuli across participants. We generated 128 unique random event sequences, one per run per participant, such that every run contained 8 unexpected and expected trials apiece, and the same number of physics (6), psychology-action (6), and psychology-environment (4) trials, and across sequences, each scenario appeared in each possible position within a trial an equal number of times.

Subject-specific functional region of interest (ssfROI) analysis

All of our primary analyses relied on the subject-specific functional region of interest (ssfROI) approach (80). The goal of this approach was to find, in individual participants, voxels that are maximally engaged during each of our hypothesized cognitive processes — social and physical prediction (identified using the DOTSloc task), early visual processing (identified using the motionLoc task), and goal-directed attention (identified using the spWMloc task) — while allowing the stereotactic location of the voxels selected to vary across people according to their unique neuroanatomy and functional organization. In Experiment 1, ssfROIs for domain-specific regions were identified using the social vs physical interaction contrasts from the DOTSloc task, ssfROIs for MD regions were identified using the unexpected > expected contrast from runs 2-4 of the VOE task, and early visual ROIs were identified using the stimuli > rest contrast from the DOTSloc task. In Experiment 2, ssfROIs for domain-specific regions were identified using the social vs physical interaction contrast from the DOTSloc task, the MD ROIs were identified using the hard > easy contrast from the spWMloc task, MT was identified using the coherent > incoherent motion contrast in the motionLoc task, and V1 was identified using the stimuli > rest contrast from the spWM task. For both experiments, for each region, for each participant, we selected the top 100 voxels (i.e., those with the highest z values) for the contrasts (listed in Figure 2B) within the corresponding fROI parcel. See SI for details about region selection and specification.

Then, we studied the responses in these 100 voxels to our primary VOE task in a set of 8 focal regions (see SI for more information about region selection and identification). For both Experiments 1-2, we focused on the unexpected and expected test events from physical scenarios and psychological scenarios involving surprising actions (physics and psychology-action; Figure 1A-B). In exploratory analyses for Experiment 2, we studied the responses of these regions to surprising physical outcomes revealed by an agent's actions (psychology-environment; Figure 5). For exploratory analyses across all domain-specific and domain-general regions, we used Bonferroni-corrected thresholds to account for the number of regions we explored (for 24 domain-general regions, $\alpha = .05/24 = .002$; for 18 domain-specific regions, $\alpha = .05/18 = .003$).

Statistical modeling of ROI data, Experiments 1-2

For our univariate focal region analysis, we modeled the average response per region as predicted by a main effect of domain, a main effect of event, and the interaction across them. Model formula: `meanbeta ~ domain * event + (1|subjectID)`. Full regression tables for all analyses are available in the SI. Our significance threshold for these analyses was $\alpha = .025$, two-tailed (correcting for 2 regions per ROI type).

Univariate region-by-region analysis

In this analysis, we took the univariate results from 18 domain-specific and 24 domain-general regions (22 for Experiment 1; excluding left and right MT), and asked whether the responses across domain-specific regions and domain-general regions are organized by domain, event, or both. The voxel selection procedure was identical to the univariate analyses, except that we selected the top 100 voxels from each region in each hemisphere (e.g. left and right APC, rather than bilateral APC), to maximize the number of regions available as input. For each region, we computed four effect sizes (Cohen's D): the magnitude of the domain effect for expected events, and separately for unexpected events (*d_domain_expected; d_domain_unexpected*), and the magnitude of the event effect for psychology-action events, and separately for physis events (*d_event_psychology; d_event_physics*). For Experiment 2, these effects were extracted from models that controlled for low-level visual features. Originally, we pre-registered this analysis over multivariate effect sizes, rather than univariate effect sizes reported here. However, due to the lack of reliable multivariate information about events, even within domains (despite clear univariate effects), we felt that we could no longer strongly interpret these results. We report the results of this pre-registered analysis in full in the SI.

We found in Experiment 1, and hypothesized and found in Experiment 2, that patterns of activity across domain-specific regions and domain-general regions will be organized more by domain than by event. To test this hypothesis, we calculated a correlation value, using a nonparametric test of independence, which uses permutation to test the null hypothesis that two vectors are statistically independent, but not assume the linearity of their dependence. For each set of regions, we calculated a correlation value relating information about domains across events, across regions (*r_domain* = cor(*d_domain_expected, d_domain_unexpected)*), and a second correlation value relating information about events across domains, across regions (*r_event* = cor(*d_event_psychology*, *d_event_physics*). To test the hypotheses that (i) *r_domain* will be larger than expected by chance, and that (ii) *r_domain* will be larger than *r_event,* we computed the bootstrapped difference between these two values under the null hypothesis (4000 iterations). The p-value was the proportion of bootstrapped observations that were equal to or greater than (i) the empirical *r_domain*, and (ii) the empirical difference between *r_domain* and *r_event*. Our significance threshold was $\alpha = .05$, one-tailed.

Acknowledgments

We gratefully acknowledge the following funding sources: DARPA Machine Common Sense Program (CW3013552), and NIH F32HD103363 (to SL). We thank: Atsushi Takahashi, Steve Shannon, and the Athinoula A. Martinos Imaging Center at the McGovern Institute at MIT for technical and administrative support; Ev Fedorenko, Jason Fischer, Caroline Robertson, Pramod RT, Kevin Smith, and Tianmin Shu for sharing data, parcels, stimuli, and task scripts; Emily Chen, Freddy Kamps, Linette Kunin, Halie Olson, and Sabrina Piccolo for help with data collection; Haoyu Du for technical assistance; Michael Cohen, Nancy Kanwisher and Josh Tenenbaum for helpful discussion; and the Saxelab, Hilary Richardson, Minjae Kim, Cambridge Writing Group, and the New PI Writing Group for feedback on an earlier draft of this paper.

References

1. T. Shu, *et al.*, AGENT: A Benchmark for Core Psychological Reasoning. *arXiv [cs.AI]* (2021).

2. K. Smith, *et al.*, “Modeling Expectation Violation in Intuitive Physics with Coarse Probabilistic Object Representations” in *Advances in Neural Information Processing Systems 32*, H. Wallach, *et al.*, Eds. (Curran Associates, Inc., 2019), pp. 8983–8993.

3. A. Needham, R. Baillargeon, Intuitions about support in 4.5-month-old infants. *Cognition* **47**, 121–148 (1993).

4. E. S. Spelke, K. Breinlinger, J. Macomber, K. Jacobson, Origins of knowledge. *Psychol. Rev.* **99**, 605–632 (1992).

5. A. L. Woodward, Infants selectively encode the goal object of an actor’s reach. *Cognition* **69**, 1–34 (1998).

6. G. Gergely, G. Csibra, Teleological reasoning in infancy: The naïve theory of rational action. *Trends Cogn. Sci.* **7**, 287–292 (2003).

7. M. Paulus, Should infant psychology rely on the violation‐of‐expectation method? Not anymore. *Infant Child Dev.* (2022).

8. A. E. Stahl, M. M. Kibbe, Great expectations: The construct validity of the violation‐of‐expectation method for studying infant cognition. *Infant Child Dev.* (2022) https:/doi.org/10.1002/icd.2359.

9. R. Baillargeon, R. M. Scott, L. Bian, Psychological Reasoning in Infancy. *Annu. Rev. Psychol.* **67**, 159–186 (2016).

10. R. Baillargeon, Physical reasoning in infancy. *The Cognitive Neurosciences* **9**, 181–204 (1995).

11. S. M. Rivera, A. Wakeley, J. Langer, The drawbridge phenomenon: representational reasoning or perceptual preference? *Dev. Psychol.* **35**, 427–435 (1999).

12. R. S. Bogartz, J. L. Shinskey, T. H. Schilling, Object permanence in five-and-a-half-month-old infants? *Infancy* **1**, 403–428 (2000).

13. Z. L. Sim, F. Xu, Another Look at Looking Time: Surprise as Rational Statistical Inference. *Top. Cogn. Sci.* (2018) https:/doi.org/10.1111/tops.12393.

14. A. E. Stahl, L. Feigenson, Violations of Core Knowledge Shape Early Learning. *Top. Cogn. Sci.* **11**, 136–153 (2019).

15. E. S. Spelke, *What Babies Know: Core Knowledge and Composition Volume 1* (Oxford University Press, 2022).

16. R. Saxe, T. Tzelnic, S. Carey, Five-month-old infants know humans are solid, like inanimate objects. *Cognition* **101**, B1–B8 (2006).

17. R. L. Frantz, J. M. Ordy, M. S. Udelf, Maturation of pattern vision in infants during the first six months. *J. Comp. Physiol. Psychol.* **55**, 907–917 (1962).

18. D. R. Peeles, D. Y. Teller, Color vision and brightness discrimination in two-month-old human infants. *Science* **189**, 1102–1103 (1975).

19. R. N. Aslin, Why Take the Cog Out of Infant Cognition? *Infancy* **1**, 463–470 (2000).

20. M. M. Haith, Who put the cog in infant cognition? Is rich interpretation too costly? *Infant Behav. Dev.* **21**, 167–179 (1998).

21. G. Raz, R. Saxe, Learning in Infancy Is Active, Endogenously Motivated, and Depends on the Prefrontal Cortices (2020) https:/doi.org/10.1146/annurev-devpsych-121318-084841 (June 24, 2021).

22. A. E. Stahl, L. Feigenson, Cognitive development. Observing the unexpected enhances infants' learning and exploration. *Science* **348**, 91–94 (2015).

23. Z. L. Sim, F. Xu, Infants preferentially approach and explore the unexpected. *Br. J. Dev. Psychol.* (2017) https:/doi.org/10.1111/bjdp.12198.

24. J. Perez, L. Feigenson, Violations of expectation trigger infants to search for explanations. *Cognition*, 104942 (2022).

25. J. Koster-Hale, *et al.*, Mentalizing regions represent distributed, continuous, and abstract dimensions of others' beliefs. *Neuroimage* **161**, 9–18 (2017).

26. L. M. DiNicola, R. M. Braga, R. L. Buckner, Parallel distributed networks dissociate episodic and social functions within the individual. *J. Neurophysiol.* **123**, 1144–1179 (2020).

27. B. Deen, K. Koldewyn, N. Kanwisher, R. Saxe, Functional Organization of Social Perception and Cognition in the Superior Temporal Sulcus. *Cereb. Cortex* **25**, 4596–4609 (2015).

28. L. Isik, K. Koldewyn, D. Beeler, N. Kanwisher, Perceiving social interactions in the posterior superior temporal sulcus. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E9145–E9152 (2017).

29. R. Saxe, D.-K. Xiao, G. Kovacs, D. I. Perrett, N. Kanwisher, A region of right posterior superior temporal sulcus responds to observed intentional actions. *Neuropsychologia* **42**, 1435–1446 (2004).

30. T. Gao, B. J. Scholl, G. McCarthy, Dissociating the detection of intentionality from animacy in the right posterior superior temporal sulcus. *J. Neurosci.* **32**, 14276–14280 (2012).

31. B. C. Vander Wyk, C. M. Hudac, E. J. Carter, D. M. Sobel, K. A. Pelphrey, Action Understanding in the Superior Temporal Sulcus Region. *Psychol. Sci.* **20**, 771–777 (2009).

32. S. Shultz, S. M. Lee, K. Pelphrey, G. McCarthy, The posterior superior temporal sulcus is sensitive to the outcome of human and non-human goal-directed actions. *Soc. Cogn. Affect. Neurosci.* **6**, 602–611 (2011).

33. R. T. Pramod, M. A. Cohen, J. B. Tenenbaum, N. Kanwisher, Invariant representation of physical stability in the human brain. *Elife* **11** (2022).

34. S. Schwettmann, J. B. Tenenbaum, N. Kanwisher, Invariant representations of mass in the human brain. *Elife* **8** (2019).

35. J. Fischer, J. G. Mikhael, J. B. Tenenbaum, N. Kanwisher, Functional neuroanatomy of intuitive physical inference. *Proc. Natl. Acad. Sci. U. S. A.* **113**, E5072–81 (2016).

36. S. Lloyd-Fox, *et al.*, Social perception in infancy: a near infrared spectroscopy study. *Child Dev.* **80**, 986–999 (2009).

37. T. Farroni, *et al.*, Infant cortex responds to other humans from shortly after birth. *Sci. Rep.* **3**, 2851 (2013).

38. T. Wilcox, J. A. Haslup, D. A. Boas, Dissociation of processing of featural and spatiotemporal information in the infant cortex. *Neuroimage* **53**, 1256–1263 (2010).

39. D. C. Hyde, C. E. Simon, F. Ting, J. I. Nikolaeva, Functional Organization of the Temporal-Parietal Junction for Theory of Mind in Preverbal Infants: A Near-Infrared Spectroscopy Study. *J. Neurosci.* **38**, 4264–4274 (2018).

40. V. Southgate, K. Begus, S. Lloyd-Fox, V. di Gangi, A. Hamilton, Goal representation in the infant brain. *Neuroimage* **85 Pt 1**, 294–301 (2014).

41. L. E. Marsh, T. L. Mullett, D. Ropar, A. F. de C. Hamilton, Responses to irrational actions in action observation and mentalising networks of the human brain. *Neuroimage* **103**, 81–90 (2014).

42. R. Ramsey, A. F. de C. Hamilton, Triangles have goals too: understanding action representation in left aIPS. *Neuropsychologia* **48**, 2773–2776 (2010).

43. B. A. Parris, G. Kuhn, G. A. Mizon, A. Benattayallah, T. L. Hodgson, Imaging the impossible: an fMRI study of impossible causal relationships in magic tricks. *Neuroimage* **45**, 1033–1039 (2009).

44. K. Friston, The free-energy principle: a unified brain theory? *Nat. Rev. Neurosci.* **11**, 127–138 (2010).

45. R. Henson, T. Shallice, R. Dolan, Neuroimaging evidence for dissociable forms of repetition priming. *Science* **287**, 1269–1272 (2000).

46. J. Jiang, C. Summerfield, T. Egner, Visual Prediction Error Spreads Across Object Features in Human Visual Cortex. *J. Neurosci.* **36**, 12746–12763 (2016).

47. L. L. Emberson, J. E. Richards, R. N. Aslin, Top-down modulation in the infant brain: Learning-induced expectations rapidly affect the sensory cortex at 6 months. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 9585–9590 (2015).

48. E. Fedorenko, J. Duncan, N. Kanwisher, Broad domain generality in focal regions of frontal and parietal cortex. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 16616–16621 (2013).

49. M. J. Kang, *et al.*, The wick in the candle of learning: epistemic curiosity activates reward circuitry and enhances memory. *Psychol. Sci.* **20**, 963–973 (2009).

50. E. Fouragnan, C. Retzler, M. G. Philiastides, Separate neural representations of prediction

error valence and surprise: Evidence from an fMRI meta-analysis. *Hum. Brain Mapp.* **39**, 2887–2906 (2018).

51. T. Nakano, H. Watanabe, F. Homae, G. Taga, Prefrontal Cortical Involvement in Young Infants' Analysis of Novelty. *Cerebral Cortex* **19**, 455–463 (2009).

52. D. M. Werchan, A. G. E. Collins, M. J. Frank, D. Amso, Role of Prefrontal Cortex in Learning and Generalizing Hierarchical Rules in 8-Month-Old Infants. *J. Neurosci.* **36**, 10314–10322 (2016).

53. C. T. Ellis, L. J. Skalaban, T. S. Yates, N. B. Turk-Browne, Attention recruits frontal cortex in human infants. *Proc. Natl. Acad. Sci. U. S. A.* **118** (2021).

54. C. E. Robertson, *et al.*, Global motion perception deficits in autism are reflected as early as primary visual cortex. *Brain* **137**, 2588–2599 (2014).

55. R. A. Poldrack, Can cognitive processes be inferred from neuroimaging data? *Trends Cogn. Sci.* **10**, 59–63 (2006).

56. E. Fedorenko, The early origins and the growing popularity of the individual-subject analytic approach in human neuroscience. *Current Opinion in Behavioral Sciences* **40**, 105–112 (2021).

57. M. Eyre, *et al.*, The Developing Human Connectome Project: typical and disrupted perinatal functional connectivity. *Brain* **144**, 2199–2213 (2021).

58. H. L. Kosakowski, *et al.*, Selective responses to faces, scenes, and bodies in the ventral visual pathway of infants. *Curr. Biol.* **0** (2021).

59. M. S. Blumberg, K. E. Adolph, Protracted development of motor cortex constrains rich interpretations of infant cognition. *Trends Cogn. Sci.* (2023) https:/doi.org/10.1016/j.tics.2022.12.014.

60. G. Gergely, Z. Nádasdy, G. Csibra, S. Bíró, Taking the intentional stance at 12 months of age. *Cognition* **56**, 165–193 (1995).

61. R. Baillargeon, E. S. Spelke, S. Wasserman, Object permanence in five-month-old infants. *Cognition* **20**, 191–208 (1985).

62. E. S. Spelke, R. Kestenbaum, D. J. Simons, D. Wein, Spatiotemporal continuity, smoothness of motion and object identity in infancy. *British Journal of Developmental Psychology* **13**, 113–142 (1995).

63. G. Csibra, S. Bíró, O. Koós, G. Gergely, One-year-old infants use teleological representations of actions productively. *Cogn. Sci.* **27**, 111–133 (2003).

64. T. D. Ullman, E. Spelke, P. Battaglia, J. B. Tenenbaum, Mind Games: Game Engines as an Architecture for Intuitive Physics. *Trends Cogn. Sci.* **21**, 649–665 (2017).

65. J. Koster-Hale, R. Saxe, Theory of mind: a neural prediction problem. *Neuron* **79**, 836–848 (2013).

66. P. Kok, L. L. F. van Lieshout, F. P. de Lange, Local expectation violations result in global

activity gain in primary visual cortex. *Sci. Rep.* **6**, 37706 (2016).

67. T. Grossmann, The development of social brain functions in infancy. *Psychol. Bull.* **141**, 1266–1287 (2015).

68. L. J. Powell, B. Deen, R. Saxe, Using individual functional channels of interest to study cortical development with fNIRS. *Dev. Sci.* (2017) https:/doi.org/10.1111/desc.12595.

69. G. Dehaene-Lambertz, E. S. Spelke, The Infancy of the Human Brain. *Neuron* **88**, 93–109 (2015).

70. J. Jara-Ettinger, H. Gweon, L. E. Schulz, J. B. Tenenbaum, The Naïve Utility Calculus: Computational Principles Underlying Commonsense Psychology. *Trends Cogn. Sci.* **20**, 589–604 (2016).

71. R. Saxe, S. Carey, N. Kanwisher, Understanding other minds: linking developmental psychology and functional neuroimaging. *Annu. Rev. Psychol.* **55**, 87–124 (2004).

72. J. Jastorff, S. Clavagnier, G. Gergely, G. A. Orban, Neural mechanisms of understanding rational actions: middle temporal gyrus activation by contextual violation. *Cereb. Cortex* **21**, 318–329 (2011).

73. M. Brass, R. M. Schmitt, S. Spengler, G. Gergely, Investigating action understanding: inferential processes versus action simulation. *Curr. Biol.* **17**, 2117–2121 (2007).

74. H. L. Kosakowski, M. A. Cohen, L. Herrara, I. Nichoson, Face-Selective Responses Present in Multiple Regions of the Human Infant Brain. *bioRxiv* (2021).

75. M. Biondi, D. A. Boas, T. Wilcox, On the other hand: Increased cortical activation to human versus mechanical hands in infants. *Neuroimage* **141**, 143–153 (2016).

76. S. Lloyd-Fox, B. Széplaki-Köllőd, J. Yin, G. Csibra, Are you talking to me? Neural activations in 6-month-old infants in response to being addressed during natural interactions. *Cortex* **70**, 35–48 (2015).

77. Y. Hakuno, M. Hata, N. Naoi, E.-I. Hoshino, Y. Minagawa, Interactive live fNIRS reveals engagement of the temporoparietal junction in response to social contingency in infants. *Neuroimage* **218**, 116901 (2020).

78. M. Biondi, A. Hirshkowitz, J. Stotler, T. Wilcox, Cortical Activation to Social and Mechanical Stimuli in the Infant Brain. *Front. Syst. Neurosci.* **15**, 510030 (2021).

79. O. Esteban, *et al.*, fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nat. Methods* **16**, 111–116 (2019).

80. E. Fedorenko, P.-J. Hsieh, A. Nieto-Castañón, S. Whitfield-Gabrieli, N. Kanwisher, New method for fMRI investigations of language: defining ROIs functionally in individual subjects. *J. Neurophysiol.* **104**, 1177–1194 (2010).

Supplemental Information: Violations of physical and psychological expectations in the human adult brain

Shari Liu, Kristen Lydic, Jerry Mei & Rebecca Saxe

For data and code required to reproduce these figures and results, see https://osf.io/sa7jy/. Please direct questions to Shari Liu, at shariliu@jhu.edu.

For additional results, see:
https://rpubs.com/shariliu/nes_exp1_univariate
https://rpubs.com/shariliu/nes_exp1_multivariate
https://rpubs.com/shariliu/nes_exp2_univariate
https://rpubs.com/shariliu/nes_exp2_multivariate

Contents:

Supplemental figures

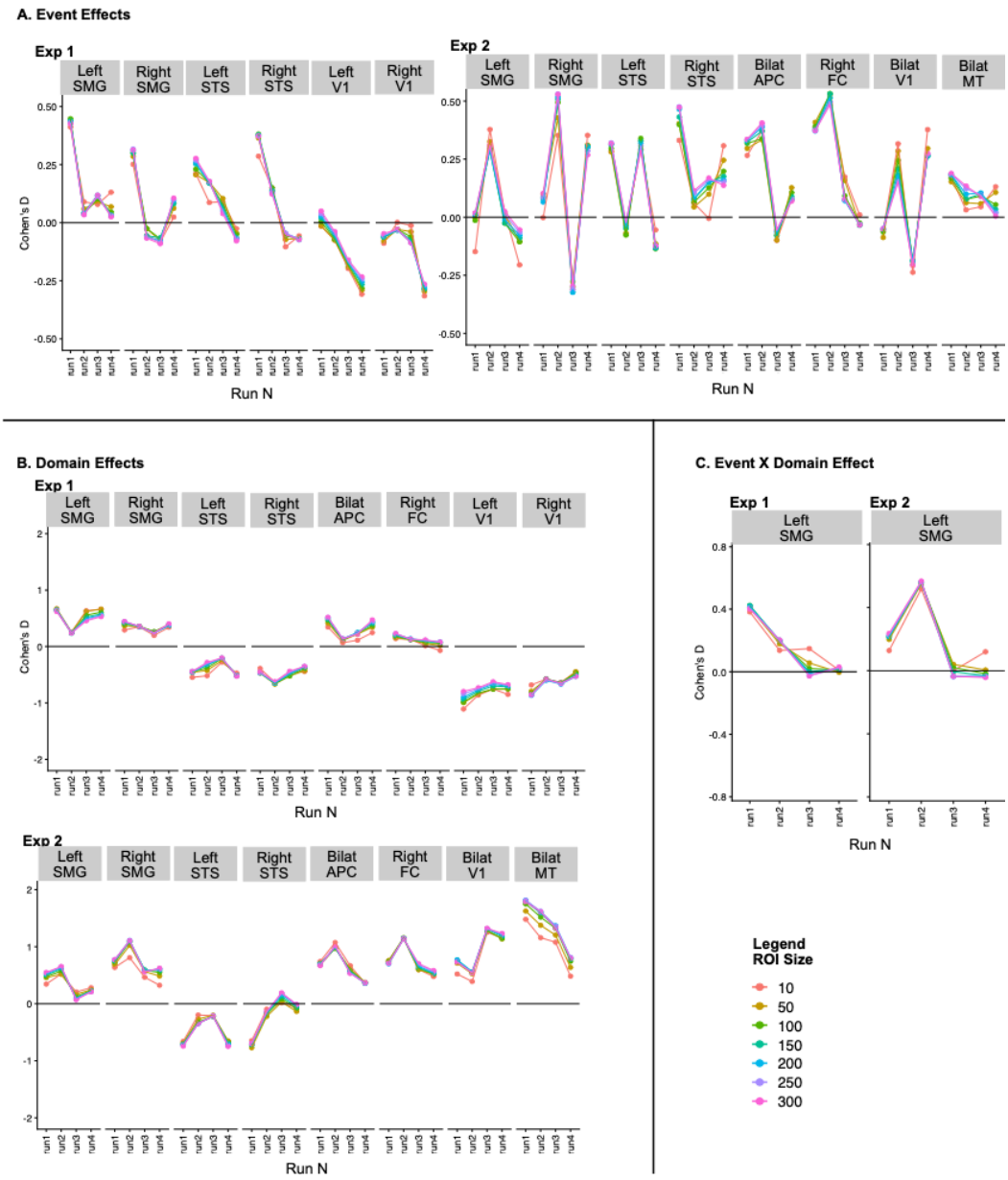


Figure S1. Sensitivity of (A) event effects (unexpected > expected), (B) the domain x event interaction (unexpected > expected, greater for physics than psychology), and (C) domain effects (physics > psychology) across experimental runs and ROI size (10-300 voxels), in the psychology-action and physics events of Experiments 1 and 2. Event effects across runs from bilateral APC and right FC are not shown for Experiment 1, because the VOE data used to choose the ROIs were from runs 2-4, and are thus non-independent from the runs 2-4 results. For all other regions, the data used to select the ROIs were independent of the data extracted from the ROIs.

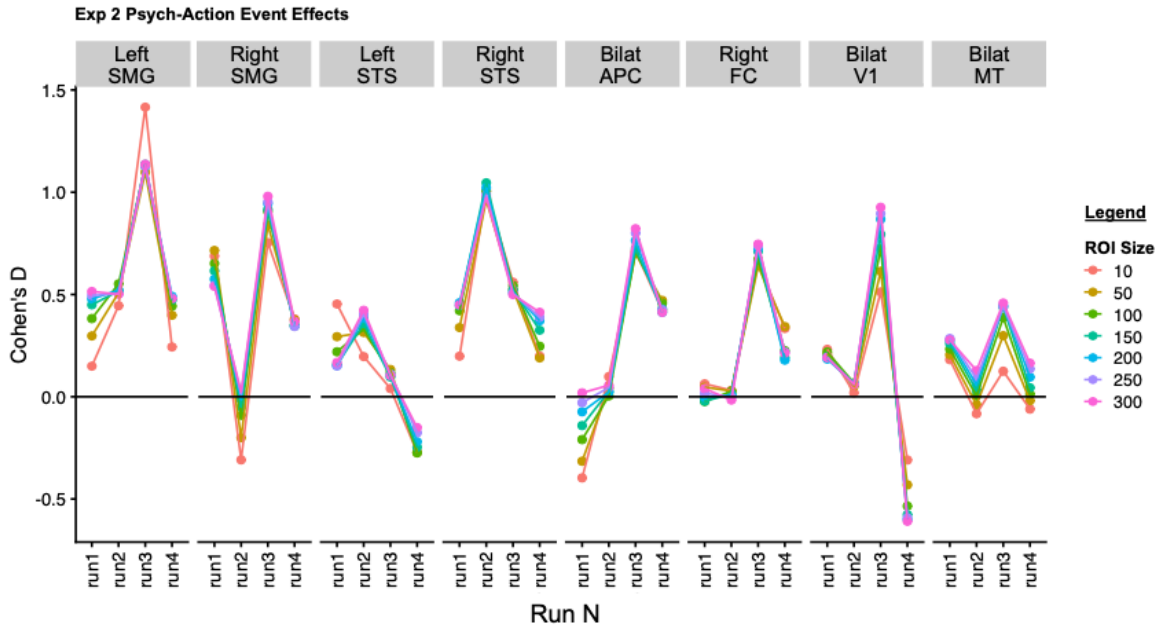


Figure S2. Sensitivity of event effects (unexpected > expected) over experimental runs and ROI size (10-300 voxels) in psychology-environment events of Experiment 2.

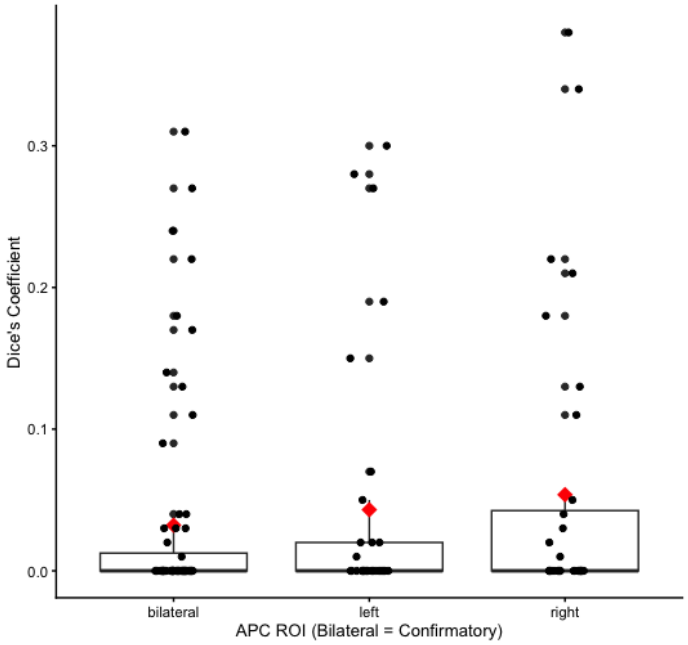


Figure S3. Dice's Coefficient (DC) between each subject's APC ROI and SMG ROIs from Experiment 2. The leftmost boxplot shows DC between bilateral APC and left and right SMG, our pre-registered ROIs. The remaining boxplots show DC between the left SMG and left APC (center), and between right SMG and right APC (right). The median DC for all three plots is 0. The mean DC is plotted in red (< .1 for all ROIs and < .05 for the pre-registered bilateral APC ROI).

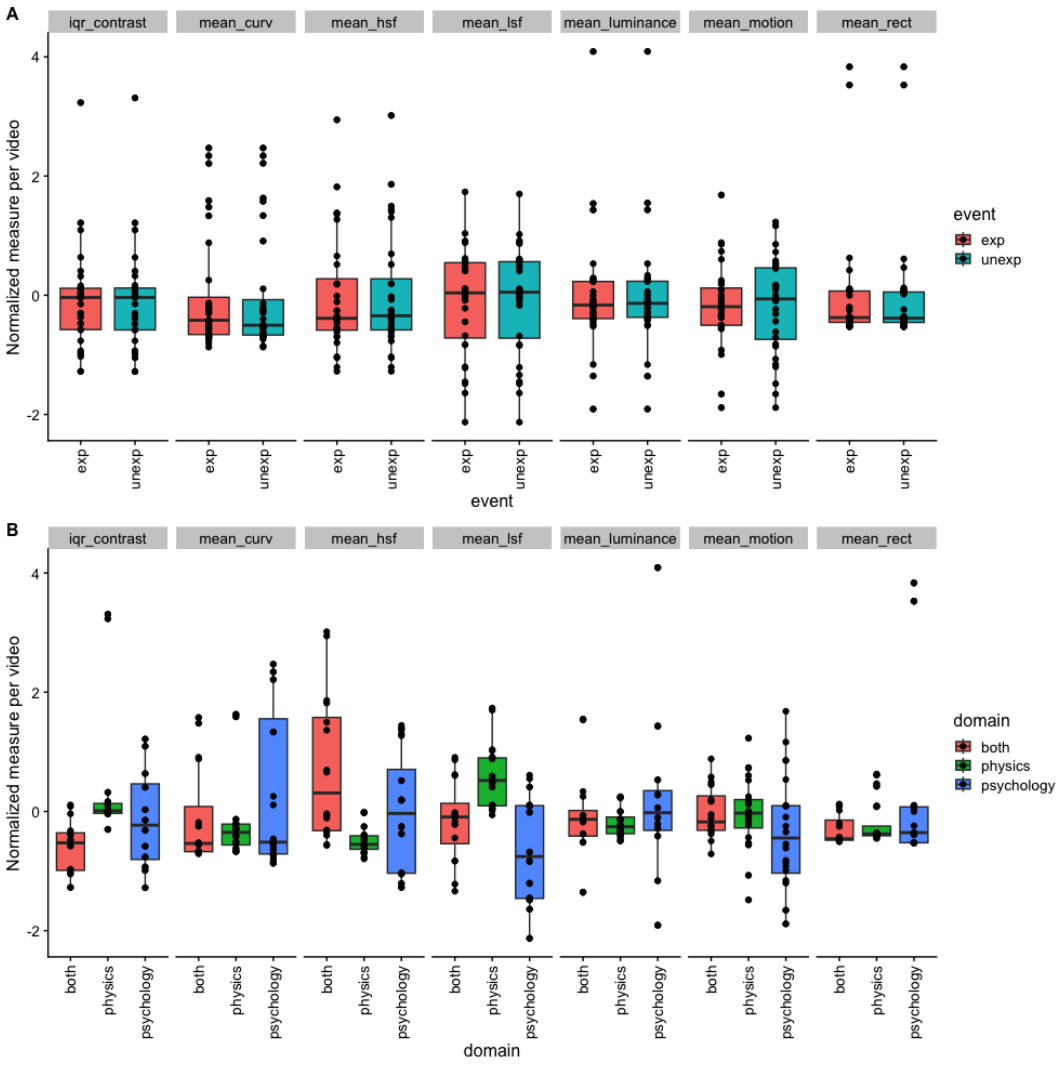
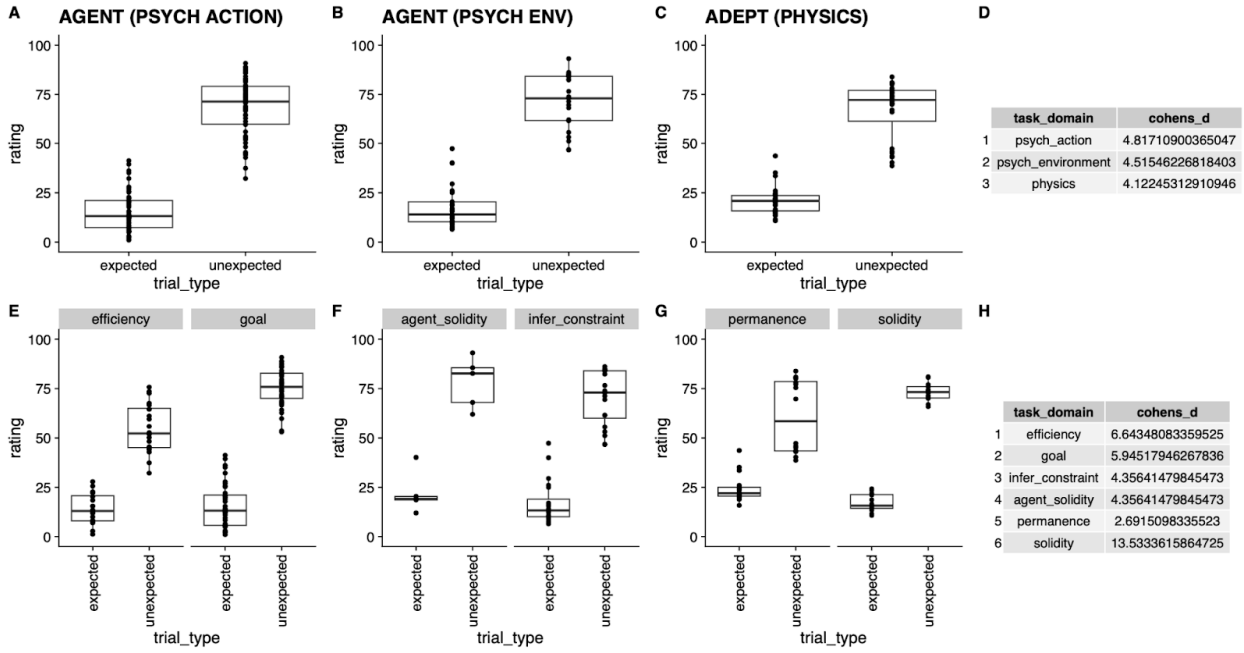


Figure S4. Boxplot of stimulus features, normalized across all videos per feature, (A) for each event type, and (B) for each domain. Each dot represents one video. Panels from left to right: stimulus contrast, curvilinearity, high spatial frequency, low spatial frequency, luminance, motion, rectilinearity.

1. Behavioral ratings for stimuli



	task_domain	cohens_d
1	psych_action	4.81710900365047
2	psych_environment	4.51546226818403
3	physics	4.12245312910946

	task_domain	cohens_d
1	efficiency	6.64348083359525
2	goal	5.94517946267836
3	infer_constraint	4.35641479845473
4	agent_solidity	4.35641479845473
5	permanence	2.6915098335523
6	solidity	13.5333615864725

Figure S5. Behavioral ratings ("How surprising?" 0: "Not at all" to 100: "Extremely") from test events in the AGENT (A-B, E-F) and ADEPT (C, G) datasets. Each dot indicates the average behavioral rating per scenario, rated by 8-10 people, grouped by domain (top row) or task (bottom row). Separate groups of people rated expected and unexpected test events from each scenario. (D-H) Effect sizes for the VOE effect (unexpected vs expected ratings), per task (H), and per domain (D).

In prior research, Smith et al (2019, ADEPT) and Shu et al. (2021, AGENT) showed adult participants (Smith et al. N = 60 total, 8 ratings per scenario; Shu et al. N = 200 total, 10 ratings per scenario) a large set of procedurally generated videos based on behavioral infant studies, a subset of which we scanned in the current paper. In these behavioral studies, adult participants saw a familiarization and test event (combined into a single event, for Smith et al.; shown as two separate events, for Shu et al), and rated how surprising each test event was on a scale of 0 to 100, with 0 indicating "not at all surprising" and 100 indicating "extremely surprising", with pairs of events presented in shuffled order, just like in our fMRI experiment.

In these behavioral studies, people from the behavioral studies never saw expected and unexpected outcomes from the same scenario. Furthermore, people saw *only* physics videos (ADEPT dataset), or *only* psychology-action and psychology-environment videos (AGENT data); these two datasets were collected separately. In our fMRI experiment, participants saw trials from both datasets, and saw both outcomes for each scenario, either immediately following each other (Exp 1), or in a separate run of the experiment (Exp 2).

Given the average ratings for each video scenario, we computed an effect size for the VOE effect (unexpected vs expected ratings) for each domain and task. We found large (d > 2) behavioral VOE effects for all tasks and domains. Notably, stimuli from the three domains were rated around equally surprising, with similar effect sizes; this suggests that any difference in neural VOE responses between domains cannot merely be explained by aggregate differences in how surprising events were across domains.

2.1 Preprocessing, Experiment 1

Results included in this manuscript come from preprocessing performed using *fMRIPprep* 1.2.6 (Esteban, Markiewicz, et al. (2018); Esteban, Blair, et al. (2018); RRID:SCR_016216), which is based on *Nipype* 1.1.7 (Gorgolewski et al. (2011); Gorgolewski et al. (2018); RRID:SCR_002502).

Anatomical data preprocessing

The T1-weighted (T1w) image was corrected for intensity non-uniformity (INU) using `N4BiasFieldCorrection` (Tustison et al. 2010, ANTs 2.2.0), and used as T1w-reference throughout the workflow. The T1w-reference was then skull-stripped using `antsBrainExtraction.sh` (ANTs 2.2.0), using OASIS as target template. Brain surfaces were reconstructed using `recon-all` (FreeSurfer 6.0.1, RRID:SCR_001847, Dale, Fischl, and Sereno 1999), and the brain mask estimated previously was refined with a custom variation of the method to reconcile ANTs-derived and FreeSurfer-derived segmentations of the cortical gray-matter of Mindboggle (RRID:SCR_002438, Klein et al. 2017). Spatial normalization to the ICBM 152 Nonlinear Asymmetrical template version 2009c (Fonov et al. 2009, RRID:SCR_008796) was performed through nonlinear registration with `antsRegistration` (ANTs 2.2.0, RRID:SCR_004757, Avants et al. 2008), using brain-extracted versions of both T1w volume and template. Brain tissue segmentation of cerebrospinal fluid (CSF), white-matter (WM) and gray-matter (GM) was performed on the brain-extracted T1w using `fast` (FSL 5.0.9, RRID:SCR_002823, Zhang, Brady, and Smith 2001).

Functional data preprocessing

For each of the 18 BOLD runs found per subject (across all tasks and sessions), the following preprocessing was performed. First, a reference volume and its skull-stripped version were generated using a custom methodology of *fMRIPrep*. The BOLD reference was then co-registered to the T1w reference using `bbregister` (FreeSurfer) which implements boundary-based registration (Greve and Fischl 2009). Co-registration was configured with nine degrees of freedom to account for distortions remaining in the BOLD reference. Head-motion parameters with respect to the BOLD reference (transformation matrices, and six corresponding rotation and translation parameters) are estimated before any spatiotemporal filtering using `mcflirt` (FSL 5.0.9, Jenkinson et al. 2002). The BOLD time-series, were resampled to surfaces on the following spaces: *fsaverage5*. The BOLD time-series (including slice-timing correction when applied) were resampled onto their original, native space by applying a single, composite transform to correct for head-motion and susceptibility distortions. These resampled BOLD time-series will be referred to as *preprocessed BOLD in original space*, or just *preprocessed BOLD*. First, a reference volume and its skull-stripped version were generated using a custom methodology of *fMRIPrep*. Automatic removal of motion artifacts using independent component analysis (ICA-AROMA, Pruim et al. 2015) was performed on the *preprocessed BOLD on MNI space* time-series after removal of non-steady state volumes and spatial smoothing with an isotropic, Gaussian kernel of 6mm FWHM (full-width half-maximum). Corresponding "non-aggresively" denoised runs were produced after such smoothing. Additionally, the "aggressive" noise-regressors were collected and placed in the corresponding confounds file. The BOLD time-series were resampled to MNI152NLin2009cAsym standard space, generating a *preprocessed BOLD run in MNI152NLin2009cAsym space*. First, a reference volume and its skull-stripped version were generated using a custom methodology of *fMRIPrep*. Several confounding time-series were calculated based on the *preprocessed BOLD*: framewise displacement (FD), DVARS and three region-wise global signals. FD and DVARS are calculated for each functional run, both using their implementations in *Nipype* (following the definitions by Power et al. 2014). The three global signals are extracted within the CSF, the WM, and the

whole-brain masks. Additionally, a set of physiological regressors were extracted to allow for component-based noise correction (*CompCor*, Behzadi et al. 2007). Principal components are estimated after high-pass filtering the *preprocessed BOLD* time-series (using a discrete cosine filter with 128s cut-off) for the two *CompCor* variants: temporal (tCompCor) and anatomical (aCompCor). Six tCompCor components are then calculated from the top 5% variable voxels within a mask covering the subcortical regions. This subcortical mask is obtained by heavily eroding the brain mask, which ensures it does not include cortical GM regions. For aCompCor, six components are calculated within the intersection of the aforementioned mask and the union of CSF and WM masks calculated in T1w space, after their projection to the native space of each functional run (using the inverse BOLD-to-T1w transformation). The head-motion estimates calculated in the correction step were also placed within the corresponding confounds file. All resamplings can be performed with *a single interpolation step* by composing all the pertinent transformations (i.e. head-motion transform matrices, susceptibility distortion correction when available, and co-registrations to anatomical and template spaces). Gridded (volumetric) resamplings were performed using `antsApplyTransforms` (ANTs), configured with Lanczos interpolation to minimize the smoothing effects of other kernels (Lanczos 1964). Non-gridded (surface) resamplings were performed using `mri_vol2surf` (FreeSurfer).

Many internal operations of *fMRIPrep* use *Nilearn* 0.5.0 (Abraham et al. 2014, RRID:SCR_001362), mostly within the functional processing workflow. For more details of the pipeline, see the section corresponding to workflows in *fMRIPrep*'s documentation.

Copyright Waiver

2.2 Preprocessing, Experiment 2

Results included in this manuscript come from preprocessing performed using fMRIPrep 22.0.2 (Esteban, Markiewicz, et al. (2018); Esteban, Blair, et al. (2018); RRID:SCR_016216), which is based on Nipype 1.8.5 (K. Gorgolewski et al. (2011); K. J. Gorgolewski et al. (2018); RRID:SCR_002502).

Anatomical data preprocessing

A total of 1 T1-weighted (T1w) images were found within the input BIDS dataset.The T1-weighted (T1w) image was corrected for intensity non-uniformity (INU) with `N4BiasFieldCorrection` (Tustison et al. 2010), distributed with ANTs 2.3.3 (Avants et al. 2008, RRID:SCR_004757), and used as T1w-reference throughout the workflow. The T1w-reference was then skull-stripped with a *Nipype* implementation of the `antsBrainExtraction.sh` workflow (from ANTs), using OASIS30ANTs as target template. Brain tissue segmentation of cerebrospinal fluid (CSF), white-matter (WM) and gray-matter (GM) was performed on the brain-extracted T1w using `fast` (FSL 6.0.5.1:57b01774, RRID:SCR_002823, Zhang, Brady, and Smith 2001). Brain surfaces were reconstructed using `recon-all` (FreeSurfer 7.2.0, RRID:SCR_001847, Dale, Fischl, and Sereno 1999), and the brain mask estimated previously was refined with a custom variation of the method to reconcile ANTs-derived and FreeSurfer-derived segmentations of the cortical gray-matter of Mindboggle (RRID:SCR_002438, Klein et al. 2017). Volume-based spatial normalization to two standard spaces (MNI152NLin2009cAsym, MNI152NLin6Asym) was performed through nonlinear registration with `antsRegistration` (ANTs 2.3.3), using brain-extracted versions of both T1w reference and the T1w template. The following templates were selected for spatial normalization: *ICBM 152 Nonlinear Asymmetrical template version 2009c* [Fonov et al. (2009), RRID:SCR_008796; TemplateFlow ID: MNI152NLin2009cAsym], *FSL's MNI ICBM 152 non-linear 6th Generation Asymmetric Average Brain Stereotaxic Registration Model* [Evans et al. (2012), RRID:SCR_002823; TemplateFlow ID: MNI152NLin6Asym].

Functional data preprocessing

For each of the 10 BOLD runs found per subject (across all tasks and sessions), the following preprocessing was performed. First, a reference volume and its skull-stripped version were generated using a custom methodology of *fMRIPrep*. Head-motion parameters with respect to the BOLD reference (transformation matrices, and six corresponding rotation and translation parameters) are estimated before any spatiotemporal filtering using `mcflirt` (FSL 6.0.5.1:57b01774, Jenkinson et al. 2002). BOLD runs were slice-time corrected to 0.95s (0.5 of slice acquisition range 0s-1.9s) using `3dTshift` from AFNI (Cox and Hyde 1997, RRID:SCR_005927). The BOLD time-series (including slice-timing correction when applied) were resampled onto their original, native space by applying the transforms to correct for head-motion. These resampled BOLD time-series will be referred to as *preprocessed BOLD in original space*, or just *preprocessed BOLD*. The BOLD reference was then co-registered to the T1w reference using `bbregister` (FreeSurfer) which implements boundary-based registration (Greve and Fischl 2009). Co-registration was configured with six degrees of freedom. Several confounding time-series were calculated based on the *preprocessed BOLD*: framewise displacement (FD), DVARS and three region-wise global signals. FD was computed using two formulations following Power (absolute sum of relative motions, Power et al. (2014)) and Jenkinson (relative root mean square displacement between affines, Jenkinson et al. (2002)). FD and DVARS are calculated for each functional run, both using their implementations in *Nipype* (following the definitions by Power et al. 2014). The three global signals are extracted within the CSF, the WM, and the whole-brain masks. Additionally, a set of physiological regressors were extracted to allow for component-based noise correction (*CompCor*, Behzadi et

al. 2007). Principal components are estimated after high-pass filtering the *preprocessed BOLD* time-series (using a discrete cosine filter with 128s cut-off) for the two *CompCor* variants: temporal (tCompCor) and anatomical (aCompCor). tCompCor components are then calculated from the top 2% variable voxels within the brain mask. For aCompCor, three probabilistic masks (CSF, WM and combined CSF+WM) are generated in anatomical space. The implementation differs from that of Behzadi et al. in that instead of eroding the masks by 2 pixels on BOLD space, a mask of pixels that likely contain a volume fraction of GM is subtracted from the aCompCor masks. This mask is obtained by dilating a GM mask extracted from the FreeSurfer's *aseg* segmentation, and it ensures components are not extracted from voxels containing a minimal fraction of GM. Finally, these masks are resampled into BOLD space and binarized by thresholding at 0.99 (as in the original implementation). Components are also calculated separately within the WM and CSF masks. For each CompCor decomposition, the *k* components with the largest singular values are retained, such that the retained components' time series are sufficient to explain 50 percent of variance across the nuisance mask (CSF, WM, combined, or temporal). The remaining components are dropped from consideration. The head-motion estimates calculated in the correction step were also placed within the corresponding confounds file. The confound time series derived from head motion estimates and global signals were expanded with the inclusion of temporal derivatives and quadratic terms for each (Satterthwaite et al. 2013). Frames that exceeded a threshold of 0.5 mm FD or 1.5 standardized DVARS were annotated as motion outliers. Additional nuisance timeseries are calculated by means of principal components analysis of the signal found within a thin band (*crown*) of voxels around the edge of the brain, as proposed by (Patriat, Reynolds, and Birn 2017). The BOLD time-series were resampled into standard space, generating a *preprocessed BOLD run in MNI152NLin2009cAsym space*. First, a reference volume and its skull-stripped version were generated using a custom methodology of *fMRIPrep*. The BOLD time-series were resampled onto the following surfaces (FreeSurfer reconstruction nomenclature): *fsaverage*. Automatic removal of motion artifacts using independent component analysis (ICA-AROMA, Pruim et al. 2015) was performed on the *preprocessed BOLD on MNI space* time-series after removal of non-steady state volumes and spatial smoothing with an isotropic, Gaussian kernel of 6mm FWHM (full-width half-maximum). Corresponding "non-aggresively" denoised runs were produced after such smoothing. Additionally, the "aggressive" noise-regressors were collected and placed in the corresponding confounds file. *Grayordinates* files (Glasser et al. 2013) containing 91k samples were also generated using the highest-resolution `fsaverage` as intermediate standardized surface space. All resamplings can be performed with *a single interpolation step* by composing all the pertinent transformations (i.e. head-motion transform matrices, susceptibility distortion correction when available, and co-registrations to anatomical and output spaces). Gridded (volumetric) resamplings were performed using `antsApplyTransforms` (ANTs), configured with Lanczos interpolation to minimize the smoothing effects of other kernels (Lanczos 1964). Non-gridded (surface) resamplings were performed using `mri_vol2surf` (FreeSurfer).

Many internal operations of *fMRIPrep* use *Nilearn* 0.9.1 (Abraham et al. 2014, RRID:SCR_001362), mostly within the functional processing workflow. For more details of the pipeline, see the section corresponding to workflows in *fMRIPrep*'s documentation.

Copyright Waiver

3. Description of data analysis pipeline

All analyses, after preprocessing, used the standard fMRI pipeline from the Saxe Lab (https://saxelab.mit.edu/).

Packages and software
We use singularity containers for all processes in the lab pipeline. We are currently working on making our containers accessible publicly, but for now here is a list of software versions we use:

- Singularity: 3.4.1
- Docker image version of singularity available here
- We also use FSL for group level analyses:
    - FSL (for randomise, cluster): 5.0.9
    - Docker image available here

Software below used within Singularity containers:

- heudiconv: 0.9.0 (data analyzed before summer 2022: 0.5.4.dev1)- singularity image from Docker here
- fmriprep: 22.0.2 (data analyzed before summer 2022: v1.2.6)- singularity image from Docker here
- nipy/nipype: 1.5.1 - singularity image from Docker (closest match) here

Conda:

- 4.5.12 (heudiconv; fmriprep)
- 4.8.4 (nipype; univariate/multivariate ROI analyses)
- Python:
    - 3.6.7 (used in heudiconv container)
    - 3.7.1 (fmriprep container)
    - 3.6.5 (nipype container)
    - 3.8.3 (univariate/multivariate ROI analyses container)

Processes
Convert DICOMs to BIDS
We use heudiconv in a singularity container to convert fMRI data to BIDS format based on the experimental design.

Preprocessing
We preprocess data using fMRI data using the fMRIprep toolbox within a singularity container. Here is an example of how we call fMRIprep using the standard flags for our pipeline:

```
fmriprep $data_directory/BIDS $data_directory/BIDS/derivatives
participant --participant_label $subject_id --mem_mb 15000 --ignore
slicetiming --use-aroma -w $scratch --fs-license-file
$FSL_license_path --output-spaces MNI152NLin6Asym:res-2
```

Note, fMRIprep is technically nondeterministic; there is slight computational variability that results in slightly different reconstructions each time fMRIprep is run. For this reason, we try to maintain a standard of sharing subject-level preprocessed data as well as raw BIDS data when possible (i.e., when we have consent to share).

fMRIprep includes standard fMRI preprocessing, and with the `--use-aroma` flag, it also runs "soft" artifact correction and generates the confounds used as nuisance regressors in first-level modeling. See fMRIprep pages (linked above) for details.

Motion exclusions
After preprocessing, we use fMRIprep's Frame Displacement (FD) estimate per run to flag volumes within each run that have greater than X units of change (typically: 0.4 units) in FD from the start of the run. These volumes are excluded from first-level analyses. If greater than Y% (typically: 25%) of any run is flagged as motion, the whole run is excluded.

First level analyses
We use Nipype to combine tools from different software packages, mainly relying on Nipype's FSL interface to fit the run-level (first-level) GLM. The model is fit using FSL's FEAT tool.

There are event regressors per each contrast specified in the study-specific contrast file, as well as confounds imported from fMRIprep preprocessing. Specifically, this step relies on the confounds text file that fmriprep outputs and the realigned and normalized bold and anatomical images, as well as the events.tsv files located inside the BIDS directory specifying the onset and duration for every condition in the experiment (instructions to create the events.tsv file below).

The design for the experiment is calculated from those event files, along with nuisance regressors specified below. Each event regressor is convolved with a double-gamma HRF, and a high-pass filter is applied to both the data and the model.

Artifact detection is performed using nipype's RapidART toolbox, which is itself an implementation of SPM's ART toolbox. Individual TRs are identified as outliers if they exceed a motion threshold of more than .4 units of frame displacement, or if the average signal intensity of that volume is more than three standard deviations away from the mean average signal intensity.

In addition to the ART outliers (one regressor per outlier volume), the current Saxelab script includes a summary movement regressor (framewise displacement, or FD), and 6 anatomical CompCor regressors that are intended to control for the average signal in white matter and CSF. All regressors other than head movement parameters were convolved with a standard double-gamma hemodynamic response function, with a high pass filter of 1/210 Hz (Experiment 1) or 1/229 Hz (Experiment 2) applied to both the data and the model. Event regressors were defined as a boxcar from the start and end of each block (localizer tasks) or event (VOE task).

A smoothing kernel of 6mm is applied to the preprocessed bold images, and finally, FSL's GLM runs the first-level model. The current default is to run the model in MNI space.

Contrasts are estimated based on the contrasts specified in the contrasts.tsv file, located in the data/BIDS/code directory.

The standard outputs of an FSL analysis are created in the output directory, including parameter estimates (pe.nii.gz), contrast estimates (con.nii.gz), and residuals. For exploring significance at the run level, the con*zstat.nii.gz are the most useful files, while higher-level models will use the cope and varcope images as inputs to their mixed-effects models.

Second level analyses
Subject-level or second-level modeling combines the GLMs across runs, per subject.

The subject-level scripts will take the data from first level analyses and do operations on them; namely, we use the copes (beta estimates) and varcopes (variance estimates) using FSL's fixed-effects flow. We again use Nipype to execute this. Specifically, we use the FSL FEAT sub-tool called FLAME (FMRIB's Local Analysis of Mixed Effects).

There are two avenues for combining run-level model outputs after creating first-level models, though the second is more commonly used, and also includes the outputs of the first:

(1) Traditional: Create a single second-level model combining all runs of a task, per subject and per contrast.

(2) Iterative: Iteratively create a second-level model for each set of n-1 runs (excluding 1 run from all n runs), per task, per subject, per contrast. (Note: we will call each of these leave-one-run-out combinations a "fold.") This allows us to e.g., select the top voxels based on n-1/n of the data and extract the betas only from the held-out run. We repeat for each possible fold (leave out each run once), then average the results from the held-out runs.

Group level analysis
During second-level modeling, we created one model for each task (VOE, DOTS, spWMloc, motionLoc) for each participant. These maps were then passed to group-level modeling, wherein for each contrast, across subjects, we used FSL's RANDOMISE to perform a nonparametric one-sample t-test of the contrast values across subjects against 0, with 5000 permutations, in MNI space, with a threshold of alpha = .05, FWE-corrected, using threshold-free cluster enhancement (TFCE). In Experiment 1, we used variance smoothing, σ=6mm, following the recommendation of (Nichols & Holmes, 2002), due to its small sample size.

4. Procedures for parcel selection and creation

4.1 Overview

We aimed to identify the neural correlates of each of the hypothesized cognitive processes that underlie the VOE effect: domain-specific psychological or physical processing, and domain-general early visual processing and/or goal-directed attention. This section describes how we chose parcels, or search spaces for subject-specific fROI definition. In total, we studied responses in 42 parcels across the cortex. We chose parcels for psychology and physics regions from a combination of prior literature, and exploratory analyses on group data from Experiment 1 that were independent of the functional data we analyzed from the VOE task (selected using runs 2-4, and used to study the responses in run 1). The 18 non-overlapping domain-specific parcels (search spaces) we created from independent data spanned regions previously implicated in theory of mind, action understanding, and physical reasoning, as well as regions in the ventral and lateral occipital cortices and parahippocampal gyrus. The 24 non-overlapping domain-general parcels came from prior work: 4 early visual regions (parcels from Pramod et al., 2022), and 20 regions from the multiple demand network (parcels from https://evlab.mit.edu/funcloc/). The early visual parcels were derived from the Desikan-Killiany and Destrieux cortical parcellations in Freesurfer, and the multiple demand parcels were created based on functional data from a probabilistic overlap map from 197 adult participants who performed a spatial working memory task (the same task we scanned, spWMloc). These regions were selected prior to data collection for Experiment 2.

Our analyses aimed to balance two considerations: to maximize sensitivity to responses in individual regions, but also to characterize the distribution of information across the cortex. In the primary exploratory (Experiment 1) and confirmatory (Experiment 2) analyses, we focused on a few regions that served as the best proxies for each hypothesized cognitive process, based on prior literature and exploratory analyses over independent data from Experiment 1. For domain-specific psychological processing, we chose left and right superior temporal sulcus (STS). For domain-specific physical processing, we chose left and right supramarginal gyrus (SMG). Both STS and SMG were chosen because in group-level analyses, these regions showed greater responses to social and physical stimuli for both the VOE and the DOTSloc tasks, in Experiment 1 (see SI Section 7). For domain-general visual processing, we chose bilateral primary visual cortex (V1) and bilateral middle temporal area (MT); because there was no independent localizer for area MT in Experiment 1, we only studied left and right V1. For domain-general goal-directed attention, we chose bilateral anterior parietal cortex (APC), and right precentral/inferior frontal cortex (RFC), based on the exploratory analyses in Experiment 1. These two MD regions, identified using runs 2-4 of the VOE task, showed the biggest VOE effect size, appeared in a meta-analysis over regions that encode reward prediction error during learning (Fouragnan et al., 2018), and are close in proximity to findings from previous research on neural responses to magic tricks (Parris et al., 2009) and curiosity inducing trivia (Kang et al., 2009). We pre-registered this selection procedure but due to an error in this analysis, we originally selected partially different MD focal regions than what is reported in this paper. For full transparency, we report the results from these regions in the SI, Section 5.6.

We pre-registered the same 8 focal regions for Experiment 2 and defined them using our localizer tasks. We took the top 100 voxels (by z statistic) that responded to physical or social events (DOTSloc; physics and psychology regions), responded more to visual stimuli than rest (spWMloc; bilateral V1), responded more to coherent than incoherent motion (MotionLoc; bilateral MT), and responded more to difficult than easy spatial working memory tasks (spWMloc; MD regions). In subsequent exploratory analyses, we studied responses in the larger set of domain-specific and domain-general regions.

4.2 Details about domain-specific parcel construction

Our domain-specific parcels were derived from group-level data on the DOTSloc task, group-level data from runs 2-4 of the VOE task (with the held-out run 1 reserved for studying the VOE response), and parcels from Pramod et al (2022) of the frontoparietal physics regions, which respond more during judgments of the physical stability of block towers than judgments of the color of the blocks in the same stimuli. First, we created *p* maps with a relaxed threshold of *p* = 0.2 (TCFE) for both the DOTSloc and VOE data, for the contrasts social >< physical. Then, we found intersecting voxels between (i) the *p* map for the physical > social contrast found intersecting voxels between the DOTSloc task and (ii) and the frontoparietal map from Pramod et al. (2022). Next, we found intersecting voxels between the *p* map for the physical > social contrast from the (i) DOTSloc task, and (ii) the VOE task. Lastly, we found intersecting voxels between the *p* map for the social > physical contrast from (i) the DOTSloc task and (ii) the VOE task. We dropped clusters that were redundant across these intersection maps or appeared in the cerebellum, flipped the parcel for left SMG over to the right hemisphere to make a right SMG parcel, and combined small clusters together. Finally, we inflated the parcels to make a generous search space, checked for intersections between parcels and removed overlapping voxels and masked the resulting parcels with an MNI brain mask for each hemisphere to ensure clean separation.

In the end, we created 4 physical clusters that were derived from an intersection of the DOTSloc and frontoparietal parcels, 4 physical clusters that were derived from an intersection of the DOTSloc and VOE tasks (physical > social), and 10 social clusters that were derived from an intersection of the DOTSloc and VOE tasks (social > physical). All of these masks were fixed before data collection in Experiment 2, and are openly available at https://osf.io/sa7jy/.

5. Additional univariate results

All univariate analyses were carried out using packages lme4 (Bates et al. 2015), lmerTest (Kuznetsova et al. 2017), and lsmeans (Lenth, 2016).

5.1 Habituation of the neural VOE signal across runs

In Experiment 1, we checked whether the size of the VOE effect declines over runs, in left and right SMG and STS, where we predicted we would find domain-specific effects. The two MD regions were excluded from this analysis because the data used to identify them, from runs 2-4, is non-independent of the data for this analysis. We fit a linear mixed effects model including the interaction between run number and event as fixed effects, and subject ID as a random intercept (formula: `meanbeta ~ extracted_run_number * event + (1|subjectID)`). We then extracted the main effect of event per run using `lsmeans()`. We found that whereas there was a significant VOE effect in run 1 (B = 0.455, p = <.001, two-tailed), this effect was absent in the other runs (run 2: B = 0.13, p = 0.247, two-tailed; run 3: B = 0.019, p = 0.865, two-tailed; run 4: B = 0.002, p = 0.989, two-tailed). Thus, we proceeded with our ssfROI data from just the first VOE run, and pre-registered this analysis procedure as a way to select between including data from all runs, or just the first 2 runs, in Experiment 2.

In Experiment 2, this event by run manipulation check was conducted in all regions for which we predicted a positive effect (left and right SMG, left and right STS, bilateral APC, right FC). Similarly to Experiment 1, we found that there were marginal or significant event effects in runs 1 and 2 (run 1: B = -0.298, p = 0.069, two-tailed; run 2: B = -0.328, p = 0.045, two-tailed), but no significant event effects in runs 3 or 4 (run 3: B = -0.054, p = 0.744, two-tailed; run 4: B = -0.066, p = 0.685, two-tailed). Thus we followed our plan to restrict all subsequent confirmatory analyses to the first two runs, using the same set of mixed effects models and significance thresholds as for Experiment 1.

For the exploratory, psychology-environment events, we again checked whether the VOE effect declined across all runs in the same regions as for the psychology-action and physics events. Unlike the VOE effects from the physics and psychology-action events, the VOE effects we explored from this stimulus set did not habituate over runs (run 1: B = 0.23, p = 0.388, two-tailed), this effect was absent in the other runs (run 2: B = 0.286, p = 0.284, two-tailed; run 3: B = 0.915, p = 0.001, two-tailed; run 4: B = 0.308, p = 0.248, two-tailed). Based on these considerations, in the main text, we presented the results from these events in the same portion of the data as our primary analysis (runs 1 and 2), and from all available data (runs 1-4). Model formula: `meanbeta ~ event + (1|subjectID)`.

See Figures S1-2 for visualizations of effect sizes across runs in both experiments.

5.2 Overlap between MD and physics ROIs

The search space for a focal multiple demand ROI, the bilateral anterior parietal cortex (APC), substantially overlapped with the search spaces for 2 physical ROIs, the left and right supramarginal gyrus (SMG). How much do the ssfROIs, which were defined for each subject based either on a working memory task (spWMloc) or a physical prediction task (DOTSloc), overlap with each other? To investigate this question, we computed Dice's Coefficient (DC; Bennett & Miller, 2010) between the APC and left/right SMG ROIs for each subject, which expresses the amount of spatial overlap between the two regions: $DC(X, Y)= 2(X \cap Y)/(X+Y)$, where X+Y is the total number of voxels across the regions X and Y (in our case, 200 voxels, 100 per ROI), and $X \cap Y$ is the number of voxels that occupy the same location. We found that the median Dice's coefficient between each of the SMG ROIs and the APC ROIs was 0 (LSMG

range: 0-0.3; RSMG range: 0-0.38). For the majority of participants (21/32 for LSMG; 20/32 for RSMG), there was no overlap between voxels most selective for physical reasoning, and those most selective for attentional demand. See Figure S3.

5.3 Non-focal region univariate results (physics and psychology-action events)
In Experiments 1 and 2, we tested for VOE effects (unexpected > expected) for physics and psychology-action events across a larger set of domain-specific parcels we made based on independent data from Experiment 1. Parcels from prior work on physical reasoning (33) and attentional demand (48) (see Methods and SI for details), using a Bonferroni corrected alpha threshold of p = .05/24 = .002 for domain-general regions (24 total), and of p = .05/18 = .003 for domain-specific regions (18 total).

Event Effects
No regions beyond our focal regions in Experiment 1 or 2 showed a main effect of event that passed these stringent significance thresholds. The one region that passed this threshold overlapped substantially with our RFC ROI. See https://rpubs.com/shariliu/nes_results, Section 5.1.1, for results from all regions.

Domain Effects
In both Experiments 1 and 2, both domain-specific and domain-general regions showed a greater response for physical than psychological events, or vice versa, that met our stringent significance threshold. See https://rpubs.com/shariliu/nes_results, Section 5.1.1, for results from all regions.

In Experiment 1, the regions that responded more to physical events were:
- Left and right visual medial cortex (physics regions): Left [0.64,1.06], B=0.85, p<.001, two-tailed, d = 0.996, BF > 1000; right [0.355,0.733], B=0.544, p<.001, two-tailed, d = 0.708, BF > 1000
- Left and right anterior parietal cortex (physics regions; both are part of our combined bilateral APC): Left [0.075,0.33], B=0.202, p=0.002, two-tailed, d = 0.391, BF = 1.114; right `r [0.075,0.33], B=0.202, p=0.002, two-tailed, d = 0.391, BF = 1.114

In Experiment 1, the regions that responded more to psychological events were:
- Left and right lateral and ventral visual cortex. Left: [-0.715, -0.369], B = -0.542, p < .001, two-tailed, d = -0.773, BF > 1000. Right: [-0.523, -0.168], B = -0.346, p < .001, two-tailed, d = -0.481, BF = 16.472

In Experiment 2, the physics ROIs with the biggest univariate preference for physical events (by effect size) were:
- Right medial visual cortex, [0.894,1.241], B=1.068, p<.001, two-tailed, d = 1.615, BF > 1000
- Left medial visual cortex, [0.704,1.11], B=0.907, p<.001, two-tailed, d = 1.174, BF > 1000
- Right superior parietal cortex, [0.671,1.175], B=0.923, p<.001, two-tailed, d = 0.962, BF > 1000

In Experiment 2, the MD/early visual ROIs with the biggest univariate preference for physical events (by effect size) were:
- Right MT, [0.611,0.877], B=0.744, p<.001, two-tailed, d = 1.468, BF > 1000
- Right posterior parietal cortex, [0.879,1.434], B=1.157, p<.001, two-tailed, d = 1.097, BF > 1000

- Right mid parietal cortex, [0.66,1.151], B=0.905, p<.001, two-tailed, d = 0.969, BF > 1000

In Experiment 2, two regions showed a greater response to psychological than physical events. Both were psychology ROIs:

- Left lateral and ventral visual cortex, [-0.768,-0.396], B=-0.582, p<.001, two-tailed, d = -0.824, BF > 1000
- Left MPFC, [-0.787,-0.275], B=-0.531, p<.001, two-tailed, d = -0.545, BF = 56.18

Event x Domain Interactions

No non-focal regions showed an event x domain interaction, even by the more lenient p < .05 threshold. See https://rpubs.com/shariliu/nes_results, Section 5.1.1, for results from all regions.

In sum, like in our confirmatory results, we found strong evidence for domain-specific responses, but weaker evidence for event-driven responses, in both ROIs that we selected and defined to be domain-specific and domain-general.

5.4 Alternative definition for psychology ROIs

Why did we fail to observe a consistent main effect of event, or a VOE effect only for psychological events, in left and right STS?

One possibility is a conceptual mismatch between the social information evoked by our independent localizer (two agents interacting socially), and the social information evoked by the VOE task (a single agent acting to achieve a physical goal, in the psychology-action events of Experiment 2). To test this possibility, we repeated the univariate analysis in left and right STS, except this time, we selected ssfROIs in STS based on the psychological > physical contrast from an independent split of the VOE task (runs 3-4 in Experiment 2; top 100 voxels by the z statistic like other analyses).

In Experiment 2, we found that left and right STS, defined based on a contrast between psychological events (involving instrumental action) and purely physical events, showed a reliable preference for psychological events (left STS: [-0.319,-0.045], B=-0.182, p=0.01, two-tailed, d = -0.35, BF = 0.317; right STS: [-0.415,-0.206], B=-0.31, p<.001, two-tailed, d = -0.78, BF = 58439.327). However, these left and right STS ROIs did not respond more to unexpected than expected events (left STS: [-0.218,0.055], B=-0.081, p=0.246, two-tailed, d = -0.156, BF = 0.022; right STS: [-0.187,0.022], B=-0.082, p=0.125, two-tailed, d = -0.207, BF = 0.027). Neither region showed an interaction between domain and event (left STS: [-0.131,0.142], B=0.005, p=0.937, two-tailed, d = 0.011, BF = 0.011; right STS: [-0.141,0.069], B=-0.036, p=0.502, two-tailed, d = -0.091, BF = 0.011).

In summary, across two ROI definitions,, we did not find evidence for domain-general or domain-specific prediction error in “psychology” STS ROIs, though STS did respond more to psychological than physical events, characteristic of its social functions.

5.5 Visual statistics

We tested for the robustness of our VOE effects (domain-specific event response in SMG, domain-general event responses in APC and RFC) accounting for the variability in the lower-level visual statistics in our stimuli. We conducted this exploratory analysis on the data from Experiment 2, which contained many more scenarios than Experiment 1, to maximize sensitivity to stimulus-driven effects.

For each video, we calculated the amount of luminance, contrast, motion, high spatial frequency info, low spatial frequency info, curvilinearity, and rectilinearity, z-scored across videos per feature.

To calculate spatial frequency, we computed Fourier transform on each frame of each video, and followed methods and cut-offs from Rajimehr et al. (2011) to calculate high and low spatial frequency per frame. To calculate rectilinearity, we applied angled Gabor filters (90° and 180°) with four different spatial frequencies (1, 2, 4, and 8) to each pixel. Averages were taken per frame of each stimulus video, and across frames per video. To calculate curvilinearity, we used a similar method using angled Gabor filters (30°, 60°, 90°, 120°, 150°, and 180°) with five different curve depths. These methods followed Kosakowski et al. (2022). To calculate luminance, we split each frame of each stimulus video into separate R, G, and B channels and computed luminance using the following formula: b = (b / 255) ** 2.2, g = (g / 255) ** 2.2, r = (r / 255) ** 2.2; luminance = 0.2126 * r + 0.7152 * g + 0.0722 * b (https://en.wikipedia.org/wiki/Relative_luminance). To calculate the contrast of each video, we converted each frame to grayscale and obtained the interquartile range of grayscale intensity. For all of the above visual features, we calculated each feature per frame and then averaged across frames to obtain a single value per stimulus. Finally, to calculate motion energy, we followed the methods of Nishimoto et al. (2011). We passed each stimulus video through a series of 3D spatiotemporal Gabor wavelet filters to determine the strength of each motion energy direction and speed. Then we calculated the mean value across all filters, resulting in one value per stimulus.

We found that two visual features, high and low spatial frequency, were highly correlated in many of the models including visual features as predictors. Low spatial frequency was excluded from all models to avoid issues of multicollinearity. See SI for full univariate results on psychology-action and physics events from Experiment 2, including these per-video features as regressors.
Then, we built a GLM to estimate one beta per presentation of each video (i.e. 16 familiarization-test pairs, 32 betas per run) and extracted these betas in the same ssfROIs as the confirmatory analysis. Finally, we fit the same mixed effects models as the confirmatory analysis, in the same regions, using these video-specific betas, while also adding in fixed effects for the 7 visual statistics. Low spatial frequency was dropped from all models given high collinearity with high spatial frequency. Model formula: `meanbeta ~ event * domain + normalized_iqr_contrast + normalized_mean_luminance + normalized_mean_motion + normalized_mean_hsf + normalized_mean_rect + normalized_mean_curv + (1|subjectID))`.

Each of the visual features predicted the amplitude of univariate activity in at least one focal region. All of the VOE results in our focal regions (both positive and negative) held taking into account these features:

- Left and right SMG did not show a main effect of event (LSMG: [-0.076, 0.209], B =0.067, p = 0.361, two-tailed, d = 0.068; RSMG: [-0.045, 0.252], B =0.103, p = 0.175, two-tailed, d = 0.101), but both regions showed an interaction between event and domain, with a greater VOE effect for physical events (LSMG: [0.063, 0.347], B =0.205, p = 0.005, two-tailed, d = 0.209; RSMG: [0.033, 0.328], B =0.18, p = 0.018, two-tailed, d = 0.177).
- Neither left nor right STS showed a main effect of event (LSTS: [-0.103, 0.257], B =0.077, p = 0.403, two-tailed, d = 0.062; RSTS: [-0.041, 0.247], B =0.103, p = 0.165, two-tailed, d = 0.103).

- Neither bilateral V1 and bilateral MT showed a main effect of event (V1: [-0.137, 0.26], B =0.061, p = 0.547, two-tailed, d = 0.045; MT: [-0.07, 0.163], B =0.047, p = 0.437, two-tailed, d = 0.058).
- Both bilateral APC and RFC showed a main effect of event, responding more to unexpected than expected events (APC: [0.084, 0.422], B =0.253, p = 0.004, two-tailed, d = 0.217; RFC: [0.084, 0.423], B =0.254, p = 0.004, two-tailed, d = 0.217).

The domain effects from all domain-specific regions held, though the SMG domain responses were weaker after controlling for visual features:

- Left and right SMG responded more to physical events (LSMG: [0.024, 0.512], B = 0.268, p = 0.032, two-tailed, d = 0.159; RSMG [0.04, 0.548], B = 0.294, p = 0.024, two-tailed, d = 0.168)
- Left and right STS responded more to psychological events (LSTS [-0.705, -0.086], B = -0.396, p = 0.013, two-tailed, d = -0.185; RSTS [-0.64, -0.139], B = -0.389, p = 0.003, two-tailed, d = -0.225)

The domain effects from MD and early visual regions were no longer statistically significant, after controlling for all visual features, except for V1.

- Neither RFC nor APC showed a main effect of domain (APC: [-0.156, 0.426], B = 0.135, p = 0.365, two-tailed, d = 0.067; RFC: [-0.006, 0.577], B = 0.286, p = 0.056, two-tailed, d = 0.142)
- MT did not show a main effect of domain (MT: [-0.01, 0.389], B = 0.19, p = 0.064, two-tailed, d = 0.137)
- V1 still responded more to physical than psychological events ([0.122, 0.813], B = 0.468, p = 0.008, two-tailed, d = 0.196).

The full results of this analysis, including effects per visual feature, and domain effects, can be found at https://rpubs.com/shariliu/nes_results in Section 5.1.5.

5.6 Results from originally selected MD ROIs

We originally pre-registered (1) bilateral insula and (2) right precentral/inferior frontal cortex as our focal MD ROIs. We discovered a mistake in this ROI definition analysis and, after fixing it, followed the same pre-registered procedure for selecting the two MD ROIs that appear in the main text. We report the results from these original two ROIs below for full transparency. See https://rpubs.com/shariliu/nes_results, Section 5.1.4, for full results.

In Experiment 1, neither bilateral IFC, nor bilateral insula, responded significantly more to unexpected than expected events (IFC: [-0.002, 0.265], B = 0.131, p = 0.056, two-tailed, d = 0.242, BF = 0.065; insula: [-0.003, 0.167], B = 0.082, p = 0.06, two-tailed, d = 0.238, BF = 0.039).

In Experiment 2, bilateral IFC, but not bilateral insula, responded significantly more to unexpected than expected events (IFC: [0.046, 0.383], B = 0.215, p = 0.013, two-tailed, d = 0.336, BF = 0.297; insula: [-0.042, 0.163], B = 0.06, p = 0.253, two-tailed, d = 0.154, BF = 0.016).

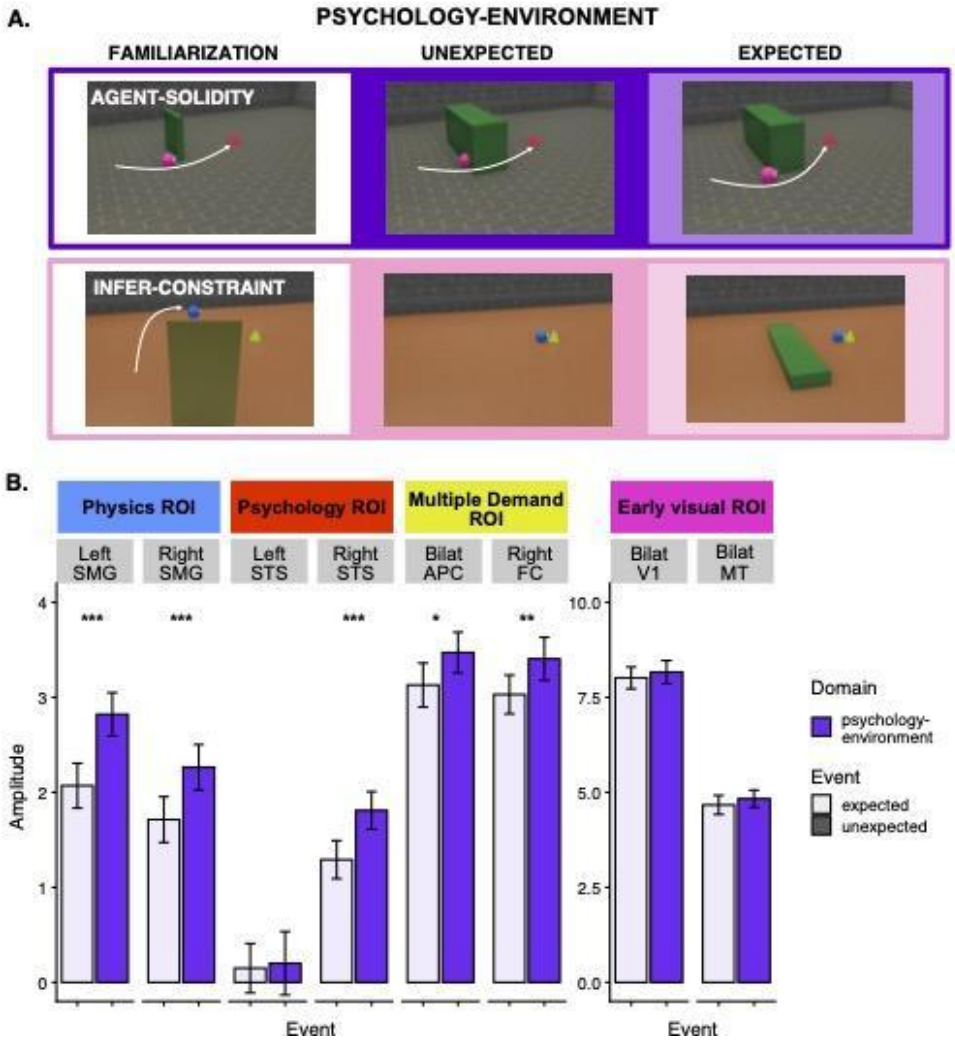


Figure S6. (A) Stimuli from the domain of intuitive psychology, wherein the actions of an agent lead to a surprising physical outcome in the surrounding environment (psychology-environment). In *agent-solidity*, an agent passes through a solid wall; in *infer-constraint*, an obstacle that explains an agent's action is missing. (B) Univariate responses towards these events in all focal regions, across all four runs. Error bars indicate within-subjects standard error.

5.7 VOE effects for psychology-environment events, all runs

Taking data from all 4 runs, we found that many non-focal regions responded significantly more to unexpected than expected psychology-environment events. See https://rpubs.com/shariliu/nes_results, Section 5.1.5, for results from all regions.
The domain-general regions:

- Left insula: [0.048, 0.167], B = 0.107, p < .001, two-tailed, d = 0.304, BF = 1.509
- Right superior frontal [0.074, 0.278], B = 0.176, p = 0.001, two-tailed, d = 0.293, BF = 1.648

The domain-specific regions:

- Right precentral/superior frontal [0.067, 0.215], B = 0.141, p < .001, two-tailed, d = 0.322, BF = 3.974
- Right superior inferior frontal [0.167, 0.353], B = 0.26, p < .001, two-tailed, d = 0.473, BF > 1000
- Left superior parietal [0.134, 0.472], B = 0.303, p < .001, two-tailed, d = 0.302, BF = 4.051
- Right superior parietal [0.157, 0.475], B = 0.316, p < .001, two-tailed, d = 0.336, BF = 15.888
- Left lateral and ventral visual [0.076, 0.275], B = 0.175, p = 0.001, two-tailed, d = 0.297, BF = 1.933
- Right lateral and ventral visual [0.127, 0.384], B = 0.256, p < .001, two-tailed, d = 0.335, BF = 12.163

5.8 VOE effects by task

We explored whether the VOE effect varied by task (e.g. permanence vs solidity), beyond by domain (e.g. physics vs psychology). We fit a mixed effects model on responses per scenario per task per ROI, extracted the coefficient and standard errors from the model, and plotted them in Figures S7-8. Because each slice of this data is small, we do not strongly interpret these results. However, qualitatively, we see that the tasks, across experiments, with the lowest neural VOE effect overall, across regions, are the permanence and infer-constraint tasks (Exp 2), and the efficiency task (Exp 1). By contrast, qualitatively, the task that evoked the highest responses across all regions was the agent-solidity task (Exp 2).

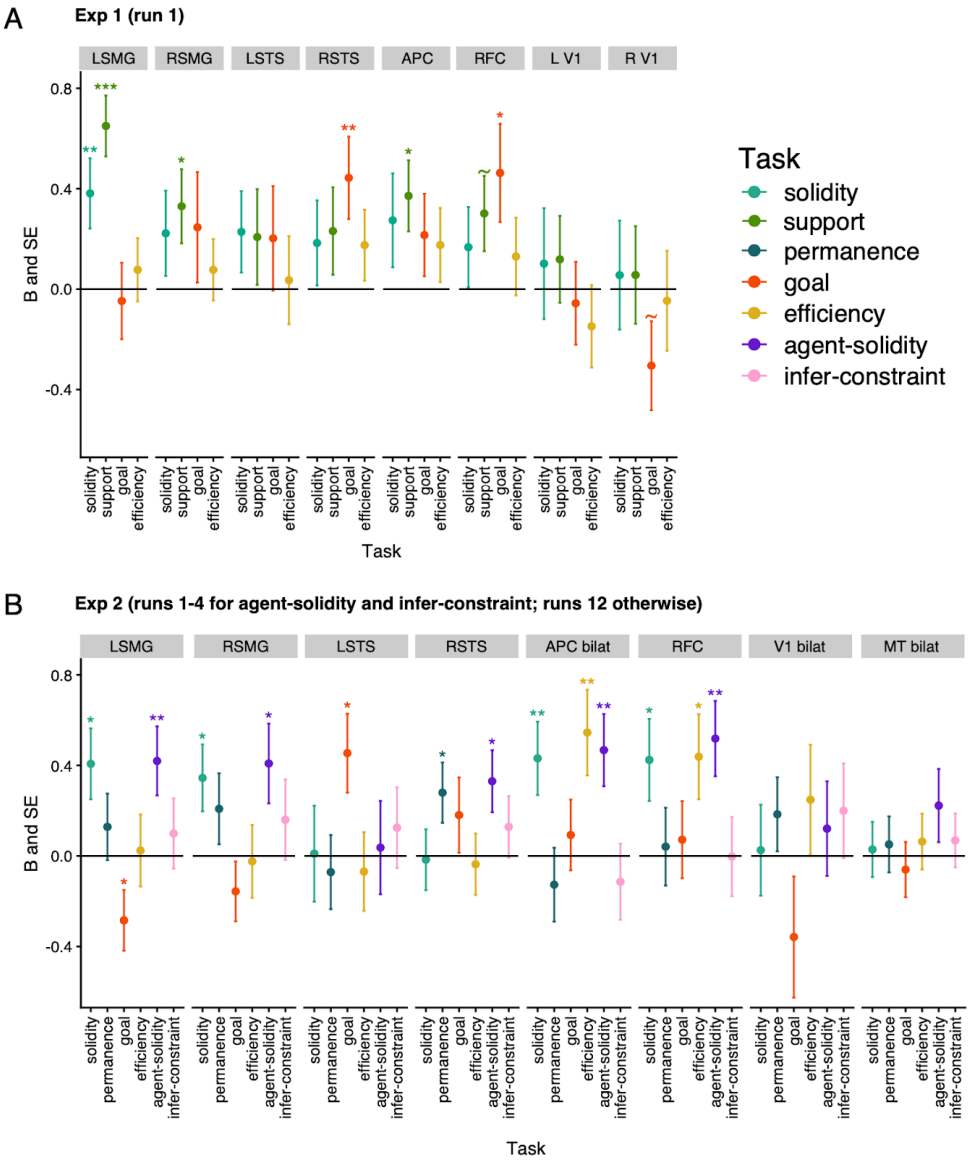


Figure S7. Per-task, per-region VOE effects for Experiments 1 and 2, organized by region. Error bars indicate the standard error of the B coefficient. ~ p < .10, * p < .05, ** p < .01, ***p < .001, two-tailed. This parallels Figure S8, except that Figure S8 is organized by task.

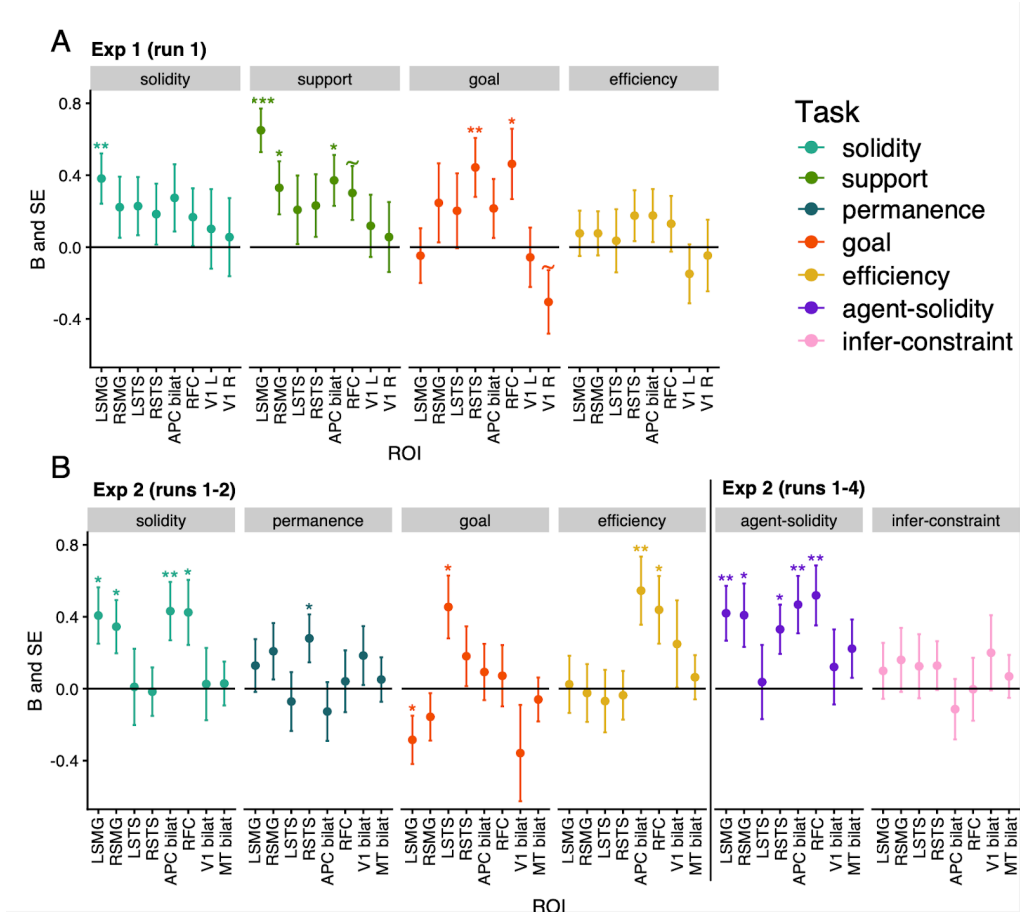


Figure S8. Estimate of neural VOE effect (positive values indicate greater responses to unexpected than expected) for each task across all focal regions for Experiments 1-2. Error bars indicate the standard error of the B coefficient. ~ p < .10, * p < .05, ** p < .01, ***p < .001, two-tailed. This parallels Figure S7, except that Figure S7 is organized by region.

6. Supplemental MVPA results

For both Experiments 1 and 2, we calculated the Euclidean distance for each participant for each region along the following category boundaries: events across domains, domains across events, events within domains (i.e. information about event within psychology-action and physics separately), and domains within events (i.e. information about domains within unexpected and expected events separately). To evaluate whether a given region had multivariate information about a given category boundary, we first computed the within vs between category distance for each boundary. Then we tested whether the within-category distances were significantly less than the between-category distances using a one-tailed Wilcoxon signed rank test. Below we will highlight the results, from Experiment 2, most relevant to our realization that we could not use MVPA to study the VOE effect from this work.

6.1 Robust univariate, and absent multivariate, event effects

In Experiment 2, both APC and RFC showed a univariate main effect of event. We planned to test for multivariate information about events that generalized across domains. However, neither of these regions contained multivariate information about event within domains (physics: $V = 257$, $p = 0.555$, one-tailed, $r = 0.104$; psychology: $V = 227$, $p = 0.756$, one-tailed, $r = 0.055$). Thus, it did not make sense to us to strongly interpret the null MVPA effect across domains (bilateral APC: $V = 273$, $p = 0.438$, one-tailed, $r = 0.137$; right frontal cortex: $V = 219$, $p = 0.8$, one-tailed, $r = 0.045$).

This stood in contrast to the robust domain multivariate effects. In Experiment 2, the regions that showed univariate domain effects, and no domain x event interaction, showed multivariate domain effects across event types:

- Bilateral V1: $V = 495$, $p < .001$, one-tailed, $r = 0.875$
- Bilateral MT: $V = 518$, $p < .001$, one-tailed, $r = 1.013$
- Bilateral APC: $V = 474$, $p < .001$, one-tailed, $r = 0.776$
- Left STS: $V = 378$, $p = 0.016$, one-tailed, $r = 0.425$
- Right STS: $V = 386$, $p = 0.011$, one-tailed, $r = 0.451$
- Right SMG: $V = 435$, $p < .001$, one-tailed, $r = 0.62$

The remaining focal region, left SMG, did not contain multivariate information about domains across events, $V = 314$, $p = 0.18$, one-tailed, $r = 0.237$. Instead, its domain boundary was marginally significant for unexpected events $V = 345$, $p = 0.067$, one-tailed, $r = 0.324$, and not significant for expected events, $V = 310$, $p = 0.2$, one-tailed, $r = 0.227$

See https://rpubs.com/shariliu/nes_results, Section 5.2, for full MVPA results for all regions, for all event boundaries, and for both experiments.

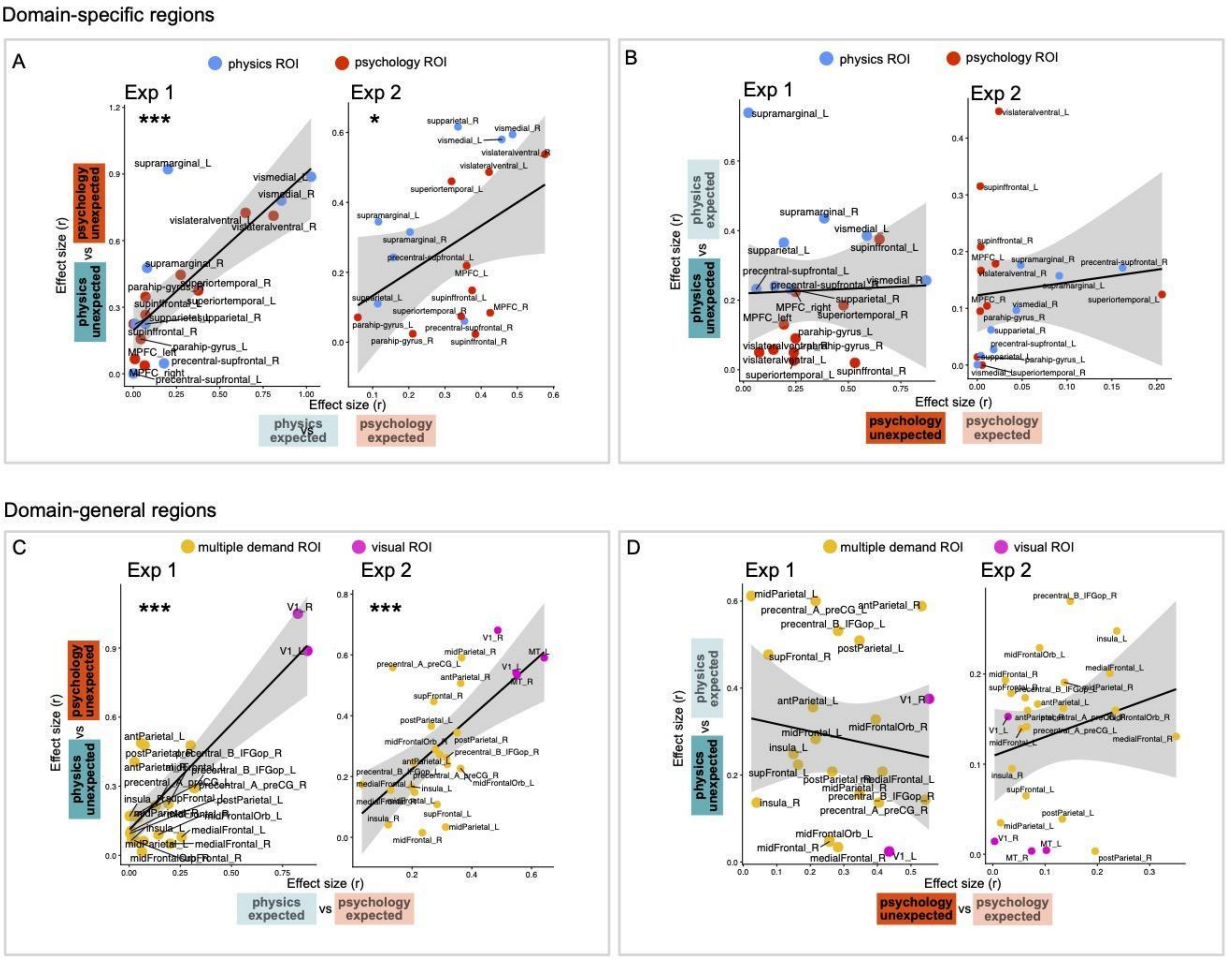


Figure S9. Multivariate effect size results across all domain-specific regions (A-B) and domain-general regions (C-D) from the exploratory results of Experiment 1, and the confirmatory results of Experiment 2. (A) and (C) show correspondence between event information across domains. (B) and (D) show correspondence between domain information across events. We found that the responses in both categories of regions were organized by domain, and not by event. ~ p < .10, * p < .05, ** p < .01, ***p < .001, one-tailed, non-parametric test for independence.

6.2 Multivariate region-by-region analysis

Here we report the results from a pre-registered confirmatory univariate analysis, studying the organization of information about domains and events across a large set of regions. However, after finding that regions that show a univariate VOE effect do not show a multivariate effect of event, within or across domains, we have decided to move that analysis, and this analysis, to the SI. Given that the multivariate information about domains is much stronger than multivariate information about events in our focal regions, we do not strongly interpret the following results, which show that both domain-specific and domain-general regions are organized more so by domain than by event. However, an alternative interpretation of these results is that the VOE response is encoded in the amplitude of activity voxels within an ROI, but the spatial pattern of this response is not reliable. Below we report them for full transparency.

For each region, we computed the effect size for the multivariate category boundary for events across and within domains, and information about domains across and within events. In Exp 1 and 2, we conducted this analysis on the 18 domain-specific regions, and on the 24 domain-general regions (22 for Experiment 1; minus left and right MT, which we had no way to define). The voxel selection procedure was identical to the univariate analyses, except that we selected the top 100 voxels from each region in each hemisphere, to maximize the number of regions available as input. In this analysis, we focused on domain-within-events and events-within-domain effect sizes - that is, how much information there is about a given category boundary in a region, relative to variance and sample size. For each region, we computed a pair of MVPA effect sizes, *r_event_psychology* and *r_event_physics*, that describes the amount of event information for each domain separately. For each region, we also computed a second pair of MVPA effect sizes, *r_domain_expected* and *r_domain_unexpected*, that describes the amount of domain information for each event type separately.

The main question is whether domain-specific regions and domain-general regions are organized primarily by domain and event, respectively. We found in Exp 1, and hypothesized and found in Exp 2, that patterns of activity across domain-specific regions and domain-general regions will be organized more by domain than by event. To test this hypothesis, we calculated a correlation value, using a nonparametric test of independence, which uses permutation to test the null hypothesis that two vectors are statistically independent, but making no assumption about the linearity of their dependence. For each set of regions, we calculated a correlation value relating information about events across domains, across regions (*r_event* = cor(*r_event_psychology*, *r_event_physics*), and a second correlation value relating information about domains across events, across regions (*r_domain* = cor(*r_domain_expected, r_domain_unexpected)*).

In Exp 2 we predicted that for both domain-specific and domain-general regions, (1) *r_domain* will be significantly larger than expected by chance (one-tailed prediction), and (2) *r_domain* will be larger than *r_event* (one-tailed prediction). To test this prediction, we used bootstrapping to compute the difference between these two values under the null hypothesis (4000 iterations). We calculated a p-value by counting the number of permuted differences out of the 4000 that was equal to or greater than the observed difference between *r_domain* and *r_event*. Our significance threshold was p = .05, one-tailed.

Exploratory results from Experiment 1 showed that for both domain-specific and domain-general regions, the degree to which a region contained information to distinguish psychological expected vs unexpected events did not significantly predict that same region's information to distinguish between physical expected and unexpected events (domain-specific: cor = -0.138, p = 0.740; domain-general: cor = -0.138, p = 0.740). In contrast, the degree to which a region contained information to distinguish between expected psychological vs physical events strongly predicted the degree to which that region distinguishes between unexpected psychological vs physical events (domain-specific: cor = 0.783, p <.001; domain-general: cor = 0.788, p < .001). Comparing the two correlations against each other using bootstrapping to generate the distribution of correlations expected under the null hypothesis (4000 iterations), we found that the domain correlation was stronger than the event correlation for domain-specific regions (95% CI [0.269, 1.065], p = 0.002), and for domain-general regions (95% CI [0.011, 1.407], p = 0.048). We then pre-registered these predictions in Experiment 2. The confirmatory analyses of Experiment 2 converged with these findings. There was no significant relationship between event information across psychological and physical events in either domain-specific regions (cor = 0.225, p = 0.148) or domain-general regions (cor = 0.225, p = 0.146). There was a correspondence between domain information across event types in both domain-specific

regions (cor = 0.439, p = 0.036) and domain-general regions (cor = 0.637, p < .001). However, the two correlations were not significantly different from each other in either domain-specific (95% CI [-0.232, 0.642], p = 0.229), nor domain-general regions (95% CI [-0.031, 0.821], p = 0.064). Taking these results literally, for both putatively domain-general and domain-specific regions, multivariate information across regions is organized by domain, and not by event. However, for the reasons described in the first paragraph of this section, we do not strongly interpret these results.

7. Whole-brain group analyses

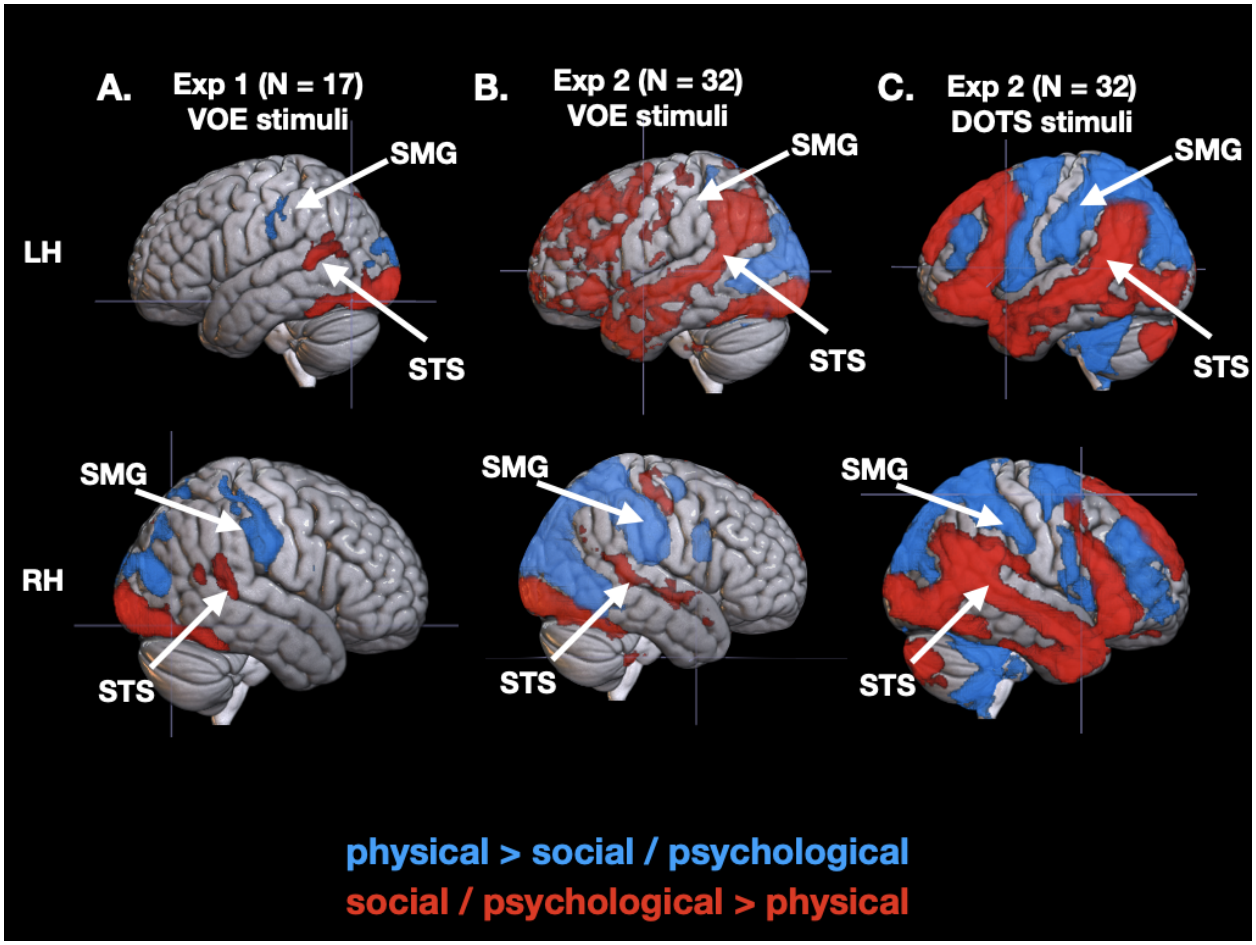


Figure S10. Results from whole-brain random effects analyses from the social/psychological vs physical contrast for the VOE task in Experiments 1-2 (A-B), and the DOTS task in Experiment 2 (C), generated from a non-parametric one-tailed test using FSL's `randomise()` and 5000 iterations, at a threshold of $p < .05$, TCFE. We additionally applied variance smoothing over the data for Experiment 1 (σ=6mm) following the recommendation of Nichols and Holmes (2002), due to the small sample size (< 20 people). Arrows point to the focal physics and psychology regions of interest (SMG and STS). Abbreviations: LH = left hemisphere; RH = right hemisphere; SMG = supramarginal gyrus; STS = superior temporal sulcus.

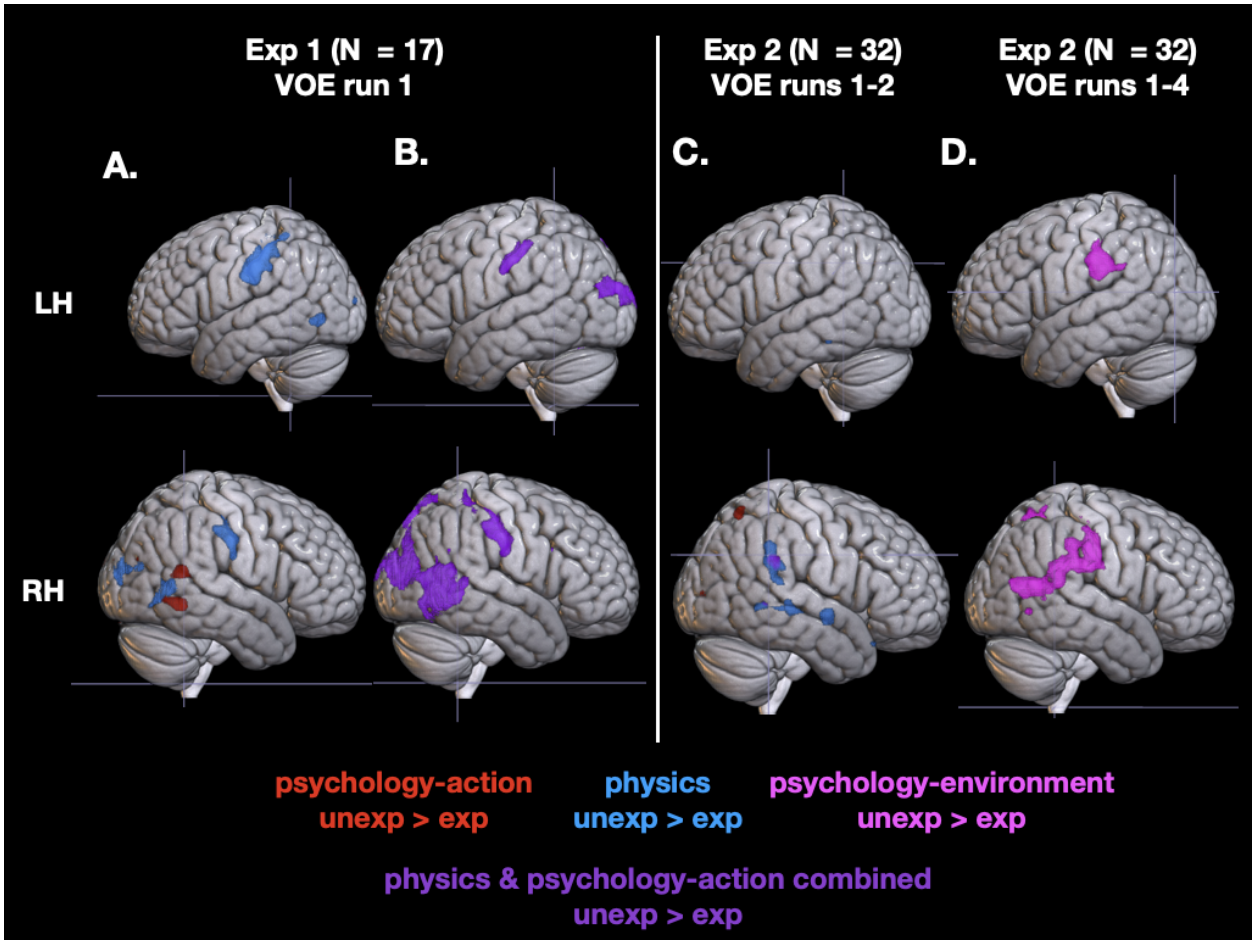


Figure S11. Group results on the neural VOE effect (unexpected > expected) from Experiment 1 (A-B) and Experiment (2) (C-D). Purple regions in (B) and (C) indicate the VOE effect folding over psychology-action and physics events, and pink regions in (D) indicate the VOE effect over psychology-environment events. All maps generated from a non-parametric one-tailed test using FSL's `randomise()` and 5000 iterations, at a threshold of $p < .05$, TCFE, except for the VOE effects for psychology-action and physics events in panel (C), which were absent at this threshold, and shown at a more lenient threshold of $p < .20$, TCFE. We additionally applied variance smoothing over the data for Experiment 1 (σ=6mm) following the recommendation of Nichols and Holmes (2002), due to the sample size (< 20 people). Abbreviations: LH = left hemisphere; RH = right hemisphere.

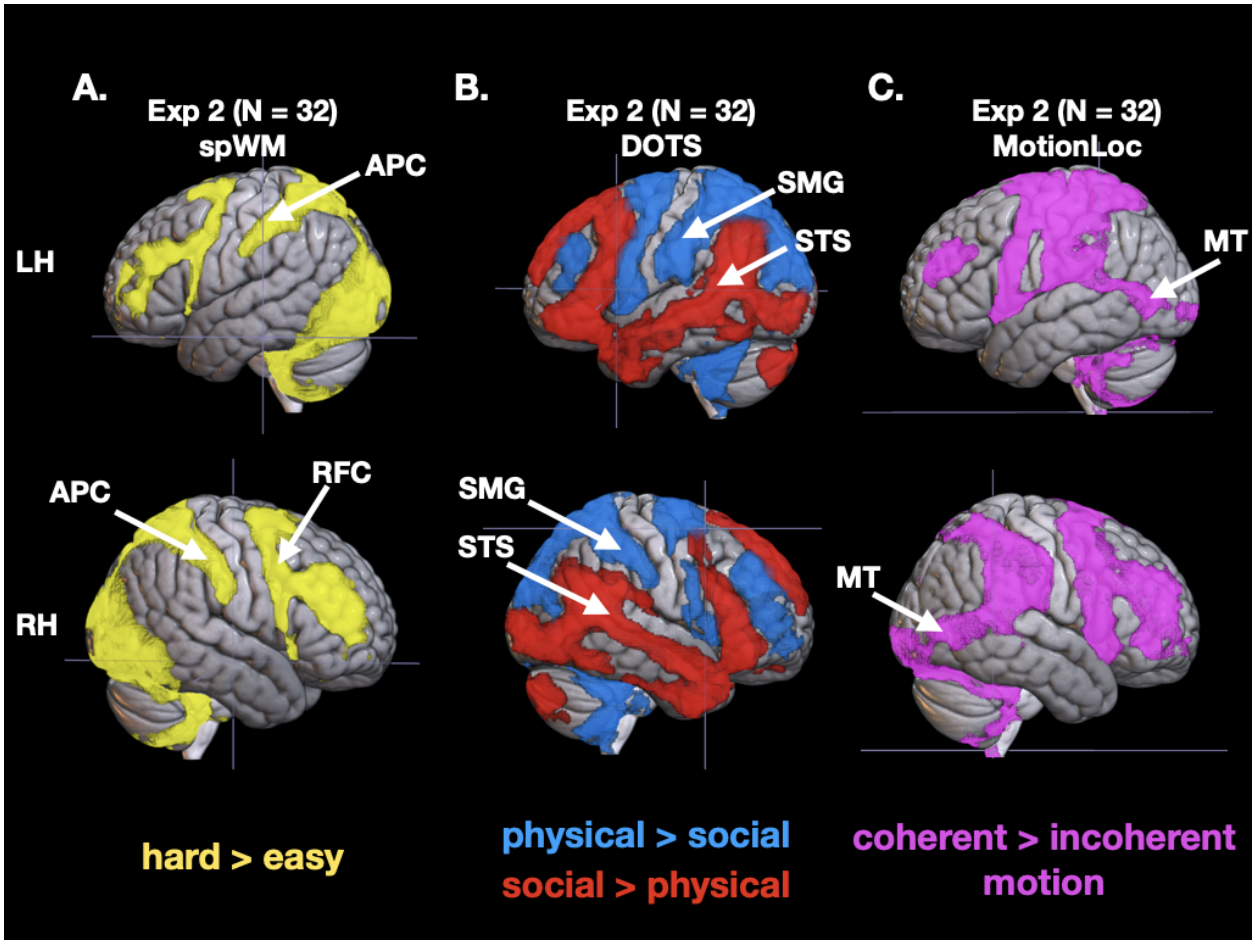


Figure S12. Group results over localizer tasks in Experiment 2. (A) Hard > easy contrast from the MD localizer. (B) Physical vs social contrast from the DOTS localizer. (C) Coherent > incoherent motion from MT localizer. These maps were generated from a non-parametric one-tailed test using FSL's `randomise()` and 5000 iterations, at a threshold of $p < .05$, TCFE. Arrows point to the focal physics and psychology regions of interest (SMG and STS), the focal MD regions of interest (APC and RFC), and one of the two focal early visual regions of interest (MT). Abbreviations: LH = left hemisphere; RH = right hemisphere; APC = anterior parietal cortex; RFC = right frontal cortex; SMG = supramarginal gyrus; STS = superior temporal sulcus; MT = motion-sensitive area.

8. References

Abraham, Alexandre, Fabian Pedregosa, Michael Eickenberg, Philippe Gervais, Andreas Mueller, Jean Kossaifi, Alexandre Gramfort, Bertrand Thirion, and Gael Varoquaux. 2014. “Machine Learning for Neuroimaging with Scikit-Learn.” *Frontiers in Neuroinformatics* 8. https://doi.org/10.3389/fninf.2014.00014.

Avants, B.B., C.L. Epstein, M. Grossman, and J.C. Gee. 2008. “Symmetric Diffeomorphic Image Registration with Cross-Correlation: Evaluating Automated Labeling of Elderly and Neurodegenerative Brain.” *Medical Image Analysis* 12 (1): 26–41. https://doi.org/10.1016/j.media.2007.06.004.

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker. "Fitting Linear Mixed-Effects Models Using lme4." *Journal of Statistical Software 67* (2015): 1-48.

Behzadi, Yashar, Khaled Restom, Joy Liau, and Thomas T. Liu. 2007. “A Component Based Noise Correction Method (CompCor) for BOLD and Perfusion Based fMRI.” *NeuroImage* 37 (1): 90–101. https://doi.org/10.1016/j.neuroimage.2007.04.042.

Bennett, Craig M., and Michael B. Miller. 2010. “How Reliable Are the Results from Functional Magnetic Resonance Imaging?” Annals of the New York Academy of Sciences 1191 (1): 133–55. https://doi.org/10.1111/j.1749-6632.2010.05446.x.

Dale, Anders M., Bruce Fischl, and Martin I. Sereno. 1999. “Cortical Surface-Based Analysis: I. Segmentation and Surface Reconstruction.” *NeuroImage* 9 (2): 179–94. https://doi.org/10.1006/nimg.1998.0395.

Esteban, Oscar, Ross Blair, Christopher J. Markiewicz, Shoshana L. Berleant, Craig Moodie, Feilong Ma, Ayse Ilkay Isik, et al. 2018. “FMRIPrep.” *Software*. Zenodo. https://doi.org/10.5281/zenodo.852659.

Esteban, Oscar, Christopher Markiewicz, Ross W Blair, Craig Moodie, Ayse Ilkay Isik, Asier Erramuzpe Aliaga, James Kent, et al. 2018. “fMRIPrep: A Robust Preprocessing Pipeline for Functional MRI.” *Nature Methods*. https://doi.org/10.1038/s41592-018-0235-4.

Evans, AC, AL Janke, DL Collins, and S Baillet. 2012. “Brain Templates and Atlases.” *NeuroImage* 62 (2): 911–22. https://doi.org/10.1016/j.neuroimage.2012.01.024.

Fedorenko, Evelina, John Duncan, and Nancy Kanwisher. 2013. “Broad Domain Generality in Focal Regions of Frontal and Parietal Cortex.” Proceedings of the National Academy of Sciences of the United States of America 110 (41): 16616–21.

Fonov, VS, AC Evans, RC McKinstry, CR Almli, and DL Collins. 2009. “Unbiased Nonlinear Average Age-Appropriate Brain Templates from Birth to Adulthood.” *NeuroImage*, Organization for human brain mapping 2009 annual meeting, 47, Supplement 1: S102. https://doi.org/10.1016/S1053-8119(09)70884-5.

Fouragnan, Elsa, Chris Retzler, and Marios G. Philiastides. "Separate neural representations of prediction error valence and surprise: Evidence from an fMRI meta‐analysis." *Human brain mapping 39*, no. 7 (2018): 2887-2906.

Glasser, Matthew F., Stamatios N. Sotiropoulos, J. Anthony Wilson, Timothy S. Coalson, Bruce Fischl, Jesper L. Andersson, Junqian Xu, et al. 2013. “The Minimal Preprocessing Pipelines for the Human Connectome Project.” *NeuroImage*, Mapping the connectome, 80: 105–24. https://doi.org/10.1016/j.neuroimage.2013.04.127.

Gorgolewski, K., C. D. Burns, C. Madison, D. Clark, Y. O. Halchenko, M. L. Waskom, and S. Ghosh. 2011. “Nipype: A Flexible, Lightweight and Extensible Neuroimaging Data Processing Framework in Python.” *Frontiers in Neuroinformatics* 5: 13. https://doi.org/10.3389/fninf.2011.00013.

Gorgolewski, Krzysztof J., Oscar Esteban, Christopher J. Markiewicz, Erik Ziegler, David Gage Ellis, Michael Philipp Notter, Dorota Jarecka, et al. 2018. “Nipype.” *Software*. Zenodo. https://doi.org/10.5281/zenodo.596855.

Greve, Douglas N, and Bruce Fischl. 2009. “Accurate and Robust Brain Image Alignment Using Boundary-Based Registration.” *NeuroImage* 48 (1): 63–72. https://doi.org/10.1016/j.neuroimage.2009.06.060.

Jenkinson, Mark, Peter Bannister, Michael Brady, and Stephen Smith. 2002. “Improved Optimization for the Robust and Accurate Linear Registration and Motion Correction of Brain Images.” *NeuroImage* 17 (2): 825–41. https://doi.org/10.1006/nimg.2002.1132.

Kang, Min Jeong, Ming Hsu, Ian M. Krajbich, George Loewenstein, Samuel M. McClure, Joseph Tao-yi Wang, and Colin F. Camerer. "The wick in the candle of learning: Epistemic curiosity activates reward circuitry and enhances memory." *Psychological science 20*, no. 8 (2009): 963-973.

Klein, Arno, Satrajit S. Ghosh, Forrest S. Bao, Joachim Giard, Yrjö Häme, Eliezer Stavsky, Noah Lee, et al. 2017. “Mindboggling Morphometry of Human Brains.” *PLOS Computational Biology* 13 (2): e1005350. https://doi.org/10.1371/journal.pcbi.1005350.

Kuznetsova, Alexandra, Per B. Brockhoff, and Rune HB Christensen. "lmerTest package: tests in linear mixed effects models." *Journal of statistical software 82* (2017): 1-26.

Kosakowski, Heather L., Michael A. Cohen, Atsushi Takahashi, Boris Keil, Nancy Kanwisher, and Rebecca Saxe. 2022. “Selective Responses to Faces, Scenes, and Bodies in the Ventral Visual Pathway of Infants.” *Current Biology: CB* 32 (2): 265-274.e5. https://doi.org/10.1016/j.cub.2021.10.064.

Lanczos, C. 1964. “Evaluation of Noisy Data.” *Journal of the Society for Industrial and Applied Mathematics Series B Numerical Analysis* 1 (1): 76–85. https://doi.org/10.1137/0701007.

Lenth, Russell V. "Least-squares means: the R package lsmeans." *Journal of statistical software 69* (2016): 1-33.

Nishimoto, Shinji, An T. Vu, Thomas Naselaris, Yuval Benjamini, Bin Yu, and Jack L. Gallant. 2011. “Reconstructing Visual Experiences from Brain Activity Evoked by Natural Movies.” *Current Biology: CB* 21 (19): 1641–46. https://doi.org/10.1016/j.cub.2011.08.031.

Parris, Ben A., Gustav Kuhn, Guy A. Mizon, Abdelmalek Benattayallah, and Tim L. Hodgson. "Imaging the impossible: An fMRI study of impossible causal relationships in magic tricks." *Neuroimage 45*, no. 3 (2009): 1033-1039.

Patriat, Rémi, Richard C. Reynolds, and Rasmus M. Birn. 2017. “An Improved Model of Motion-Related Signal Changes in fMRI.” *NeuroImage* 144, Part A (January): 74–82. https://doi.org/10.1016/j.neuroimage.2016.08.051.

Pramod, R. T., Michael A. Cohen, Joshua B. Tenenbaum, and Nancy Kanwisher. 2022. “Invariant Representation of Physical Stability in the Human Brain.” *eLife* 11 (May). https://doi.org/10.7554/eLife.71736.

Power, Jonathan D., Anish Mitra, Timothy O. Laumann, Abraham Z. Snyder, Bradley L. Schlaggar, and Steven E. Petersen. 2014. “Methods to Detect, Characterize, and Remove Motion Artifact in Resting State fMRI.” *NeuroImage* 84 (Supplement C): 320–41. https://doi.org/10.1016/j.neuroimage.2013.08.048.

Pruim, Raimon H. R., Maarten Mennes, Daan van Rooij, Alberto Llera, Jan K. Buitelaar, and Christian F. Beckmann. 2015. “ICA-AROMA: A Robust ICA-Based Strategy for Removing Motion Artifacts from fMRI Data.” *NeuroImage* 112 (Supplement C): 267–77. https://doi.org/10.1016/j.neuroimage.2015.02.064.

Rajimehr, R., Devaney, K. J., Bilenko, N. Y., Young, J. C., & Tootell, R. B. H. (2011). The “parahippocampal place area” responds preferentially to high spatial frequencies in humans and monkeys. PLoS Biology, 9(4), e1000608. https://doi.org/10.1371/journal.pbio.1000608

Satterthwaite, Theodore D., Mark A. Elliott, Raphael T. Gerraty, Kosha Ruparel, James Loughead, Monica E. Calkins, Simon B. Eickhoff, et al. 2013. “An improved framework for confound regression and filtering for control of motion artifact in the preprocessing of

resting-state functional connectivity data." *NeuroImage* 64 (1): 240–56. https://doi.org/10.1016/j.neuroimage.2012.08.052.

Shu, Tianmin, Abhishek Bhandwaldar, Chuang Gan, Kevin Smith, Shari Liu, Dan Gutfreund, Elizabeth Spelke, Joshua Tenenbaum, and Tomer Ullman. 18--24 Jul 2021. "AGENT: A Benchmark for Core Psychological Reasoning." In Proceedings of the 38th International Conference on Machine Learning, edited by Marina Meila and Tong Zhang, 139:9614–25. Proceedings of Machine Learning Research. PMLR.

Smith, Kevin, Lingjie Mei, Shunyu Yao, Jiajun Wu, Elizabeth Spelke, Josh Tenenbaum, and Tomer Ullman. 2019. "Modeling Expectation Violation in Intuitive Physics with Coarse Probabilistic Object Representations." Advances in Neural Information Processing Systems 32.

Tustison, N. J., B. B. Avants, P. A. Cook, Y. Zheng, A. Egan, P. A. Yushkevich, and J. C. Gee. 2010. "N4ITK: Improved N3 Bias Correction." *IEEE Transactions on Medical Imaging* 29 (6): 1310–20. https://doi.org/10.1109/TMI.2010.2046908.

Zhang, Y., M. Brady, and S. Smith. 2001. "Segmentation of Brain MR Images Through a Hidden Markov Random Field Model and the Expectation-Maximization Algorithm." *IEEE Transactions on Medical Imaging* 20 (1): 45–57. https://doi.org/10.1109/42.906424.