Violations of physical and psychological expectations in the human adult brain

Shari Liu [1,2] (shariliu@jhu.edu)
Kirsten Lydic [2] (kolydic@mit.edu)
Jerry Lingjie Mei [3] (lm5483@princeton.edu)
Rebecca Saxe [2] (saxe@mit.edu)

[1] Dept. Psychological and Brain Sciences, Johns Hopkins University Baltimore, MD, USA
[2] Dept. Brain and Cognitive Sciences, MIT, Cambridge, MA, USA
[3] Dept Computer Science, Princeton University, Princeton, NJ, USA

Abstract

When adults see one solid object pass through another, or see a person take the long route to a destination when a shortcut was available, we classify those events as surprising. Infants look infants look longer at the same unexpected outcomes, compared with visually similar but expected outcomes, in violation-of-expectation (VOE) experiments. What domain-specific and domain-general cognitive processes support these judgments? In a pre-registered experiment, we scanned 32 adults using functional magnetic resonance imaging (fMRI) while they watched videos designed for infant research. One region implicated in physical reasoning responded selectively to unexpected physical events, providing evidence for domain-specific physical prediction error. Multiple demand regions responded more to unexpected events regardless of domain, providing evidence for domain-general goal-directed attention. Early visual regions responded equally to unexpected and expected events, providing evidence against stimulus-driven prediction error. Thus, in adults, VOE involves domain-specific, and high-level, domain-general computations.

Keywords: cognitive neuroscience; cognitive development; intuitive psychology; intuitive physics

Introduction

In the first year of life, babies rapidly acquire expectations about the properties and behavior of inanimate objects, and animate agents. Like adults, they distinguish between surprising events (e.g. when a ball rolls off the edge of a table, and hovers in midair) and visually similar but unsurprising events (e.g. when the ball stops rolling before it reaches the edge of the table), looking longer at the unexpected outcome. Babies show this so-called violation-of-expectation (VOE) response towards events that adults rate as surprising (Shu et al., 2021; Smith et al., 2019): When objects float in midair (Needham & Baillargeon, 1993), or appear to pass through each other (Spelke, Breinlinger, Macomber, & Jacobson, 1992), and when agents change their goals (Woodward, 1998), or act inefficiently (Gergely & Csibra, 2003). However, the cognitive processes that drive longer looking in these studies are still hotly debated (Aslin, 2007; Paulus, 2022; Stahl & Kibbe, 2022). Do babies truly have domain-specific expectations about the psychological and physical world (Carey, 2009; Spelke, 2022)? Are there stimulus-driven alternative explanations that could also explain these patterns of looking (Rivera, Wakeley, & Langer, 1999)? And do babies look longer because they have detected the surprising outcome, or also because they are motivated to explore and explain the source of surprise (Sim & Xu, 2018; Stahl & Feigenson, 2019)?

Both adults' judgments of surprise, and infants' looking responses, are generated from a complex set of mental processes that are opaque to measurement, making it difficult to test these alternative hypotheses using behavior alone. However, studying the neural correlates of this behavior can shed light on this question by identifying and characterizing the domain-general and domain-specific brain networks involved in processing these events. Thus in the current research, we scanned the brains of human adults while they watched events that were designed to test for physical and psychological expectations in infants. We systematically test the (not mutually-exclusive) hypotheses that domain-specific processes, like psychological and physical reasoning, and domain-general processes, like early visual processing and goal-directed attention, underlie the violation-of-expectation response by studying the responses in regions associated with each process to these events. Broadly, domain-general hypotheses (H1) predict greater neural responses to unexpected than expected events that generalize across the domains of psychology and physics. Domain-specific hypotheses (H2) predict greater responses to unexpected events in different regions, depending on the domain.

Studying adult brains, rather than infant brains, allows us to define regions involved in each hypothesized process using independent tasks, and thus gives us more confidence that the responses we measure actually correspond to the hypothesized mental processes, without needing to rely on strong reverse inference given neural activity alone. Because of the correspondence between the large-scale topography of brain networks and evoked functional responses between adults and infants, as early as they can be measured (Dehaene-Lambertz & Spelke, 2015; Eyre et al., 2021), it is plausible that the insights we learn from adult brains could be used to guide studies of infants in future work. In the following sections, we motivate the evidence for each hypothesis from the developmental literature, and also discuss evidence from cognitive neuroscience about the likely neural correlates of each process.

Domain-general hypotheses

The first 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

In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society*.

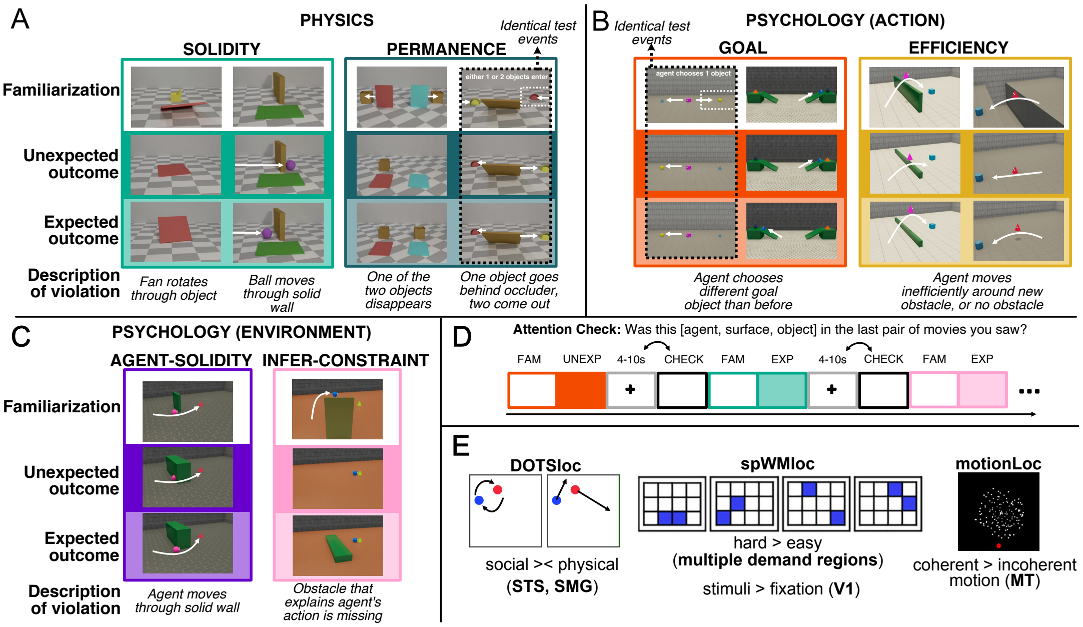
A
PHYSICS
SOLIDITY
PERMANENCE
Identical test events
either 1 or 2 objects enter
Familiarization
Unexpected outcome
Expected outcome
Description of violation
Fan rotates through object
Ball moves through solid wall
One of the two objects disappears
One object goes behind occluder, two come out
B
Identical test events
PSYCHOLOGY (ACTION)
GOAL
EFFICIENCY
agent chooses 1 object
Agent chooses different goal object than before
Agent moves inefficiently around new obstacle, or no obstacle
C
PSYCHOLOGY (ENVIRONMENT)
AGENT-SOLIDITY
INFER-CONSTRAINT
Familiarization
Unexpected outcome
Expected outcome
Description of violation
Agent moves through solid wall
Obstacle that explains agent's action is missing
D
Attention Check: Was this [agent, surface, object] in the last pair of movies you saw?
FAM
UNEXP
4-10s
CHECK
FAM
EXP
4-10s
CHECK
FAM
EXP
+
+
...
E
DOTSloc
social >< physical
(STS, SMG)
spWMloc
hard > easy
(multiple demand regions)
stimuli > fixation (V1)
motionLoc
coherent > incoherent motion (MT)

Figure 1: Overview of our violation-of-expectation (VOE) task. (A) Stimuli from the domain of intuitive physics, including violations of object **solidity** and **permanence**. (B) Stimuli from the domain of intuitive 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**). (C) Stimuli from the domain of intuitive psychology, that also involve agents acting in goal-directed ways; however here, the source of the violation is the 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. (D) Experimental design, with each trial containing a familiarization movie followed by either an expected or unexpected movie, an attention check, and then a jittered fixation period. Half of the subjects saw the attention check before the fixation cross within a trial; the other half of the subjects saw the attention check afterwards. (E) Localizer tasks and contrasts for physics and psychology regions (DOTSloc), multiple demand regions (spWMloc), and early visual regions (motionLoc).

of the unexpected stimulus). While the infant studies under discussion do account for some simple perceptual explanations (e.g. infants look longer when an agent takes an inefficient vs efficient path towards a goal; but look equally at the same paths of motion if the agent acted inefficiently to begin with; Gergely, Nádasdy, Csibra, & Bíró, 1995), the field remains divided about whether the stimuli of interest contain confounding low-level visual features (Heyes, 2014; Sirois & Jackson, 2012). After all, the behavioral methods used to study infant cognition were originally invented to study visual perception (Frantz, Ordy, & Udelf, 1962; Peeles & Teller, 1975). Research in cognitive neuroscience on neural habituation shows that new visual stimuli, relative to repeated visual stimuli, evoke activity in early visual regions and downstream in the inferior temporal cortex, in both adults and babies (Jiang, Summerfield, & Egner, 2016; Henson, Shallice, & Dolan, 2000; Emberson, Boldin, Robertson, Cannon, & Aslin, 2018).

Unexpected events may also evoke domain-general *curiosity and motivation* to gain information about the source of surprise (Sim & Xu, 2018; Stahl & Feigenson, 2019). After viewing an unexpected physical event, such as a ball rolling through a solid wall, babies show enhanced learning about that object (Stahl & Feigenson, 2015), and choose to explore that object (Sim & Xu, 2017) as though they are trying to explain the outcome (e.g. by banging the ball) (Stahl & Feigenson, 2015). In addition, longer looking in these experiments can be 'explained away': Babies look longer when a ball passes through a solid wall, rather than stopping short of the wall, but not when they first see that the wall has an archway through it, allowing the ball to pass through (Perez & Feigenson, 2022). From cognitive neuroscience, regions within the multiple demand (MD) network (Fedorenko, Duncan, & Kanwisher, 2013), including portions of the frontal and parietal cortices, plausibly support endogenous goal-directed attention. These regions 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. arithmetic vs motor inhibition). Portions of the lateral frontal cortices, in human infants, respond with greater activity when they hear or see stimuli that violate a previously learned pattern (Nakano, Watanabe, Homae, & Taga, 2009; Werchan, Collins, Frank, & Amso, 2016). In adults, these regions are engaged when people consider curiosity-inducing trivia questions, watch magic tricks, and monitor the outcomes of uncertain gambles (Kang et al., 2009; Parris, Kuhn, Mizon, Benattayallah, & Hodgson, 2009; van Lieshout, Vandenbroucke, Müller, Cools, & de Lange, 2018).

Domain-specific hypotheses

The second broad hypothesis under consideration is that surprising events violate *distinctly physical and psychological expectations*: that objects are solid and permanent entities, and that agents have goals and do not tend to change them abruptly or pursue them in unnecessarily costly ways. A strong interpretation of the developmental literature is that babies have 'core knowledge': an early-emerging conceptual repertoire consisting of domain-specific modules for different domains of thought, including physics, psychology, number, and space (Spelke, 2022). In cognitive neuroscience, there is evidence that distinct sets of 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) represents psychological information including other people's mental states and social interactions (Deen, Koldewyn, Kanwisher, & Saxe, 2015; Isik, Koldewyn, Beeler, & Kanwisher, 2017; Saxe, Xiao, Kovacs, Perrett, & Kanwisher, 2004; DiNicola, Braga, & Buckner, 2020). A second set of regions including supplementary motor area, superior parietal cortex, and supramarginal gyrus (SMG), represents physical information including object mass and stability (Pramod, Cohen, Tenenbaum, & Kanwisher, 2022; Schwettmann, Tenenbaum, & Kanwisher, 2019; Fischer, Mikhael, Tenenbaum, & Kanwisher, 2016). In infants, regions in the temporal and parietal cortices are similarly implicated in social and physical processing, respectively (Lloyd-Fox et al., 2009; Farroni et al., 2013; Wilcox, Haslup, & Boas, 2010).

Overview of current research

In the current research, we identified regions involved in early visual processing, goal-directed attention, and physical and psychological reasoning (see Figure 2A) in individual participants using validated tasks from prior literature (Fischer et al., 2016; Fedorenko et al., 2013; Robertson et al., 2014). We then measured the responses of these regions to unexpected and expected psychological and physical videos designed for infant studies. We tested whether the responses in each region are driven by manipulations of domain (psychology versus physics), event (expected versus unexpected), or a selective interaction of these factors (e.g. unexpected versus expected, only for physical events). Based on a previous exploratory study using similar materials, we pre-registered a complete analysis plan for these data. Here we report the results of the confirmatory hypothesis tests for the univariate (i.e. average magnitude of) responses in all tested regions.

Method

Open Science Practices

The methods and analyses of this experiment were pre-registered prior to data collection at `https://osf.io/8ywah`. All experiment scripts, including stimuli shown to participants, as well as the data and analysis scripts required to reproduce these results, can be found at `https://osf.io/sa7jy/`.

Participants

We recruited 33 participants (Mean age = 25.7y, range 18-45; 30 right-handed; 21 identifying as female, 12 identifying as male; 19 identifying as White, 14 identifying as Black, Asian, or biracial) from the [redacted] area. One participant withdrew from the experiment without contributing usable functional data, and was excluded from our sample, leaving a final *N* of 32. Participants had normal or corrected-to-normal vision, and no MRI contraindications. We chose this sample size using a combination of simulation power analyses over a prior experiment of *N*=17, and other considerations of time and cost. All study procedures were approved by [redacted]. Participants were asked to provide written informed consent prior to participation, and were paid $30 per hour.

Localizer tasks

Social versus physical interaction (DOTSloc) The DOTSloc task (Fischer et al., 2016) reliably evokes responses in the superior temporal sulcus (STS) and supramarginal gyrus (SMG) (our a priori regions of interest for psychological and physical reasoning). 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. Each run included 19 blocks (8 physical blocks, 8 social blocks, 3 rest blocks). On social and physical blocks, participants saw two different videos from the corresponding condition. Each run lasted approximately 8.2 minutes.

Spatial working memory (spWMloc) The spWMloc task (Fedorenko et al., 2013) identifies regions in the multiple demand (MD) network, including the inferior frontal gyrus (IFG) and the insula (our regions of interest 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 button press which of two alternative grids matched the resulting layout, with feedback. Each run included 20 16-second blocks (6 easy, 6 hard, and 4 rest blocks), and lasted approximately 7.5 minutes.

Motion (motionLoc) The motionLoc task (Robertson et al., 2014) identifies motion-sensitive regions (MT) (one region of interest for early visual processing). This task contrasts coherent vs random dot motion to enable 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 uniform direction) or randomly around the space. Participants pressed a button whenever the red dot flickered. Each run lasted approximately 4.6 minutes.

Primary task (VOE)

Stimulus design Our violation-of-expectation (VOE) stimuli were selected from 2 large-scale procedurally generated video datasets, inspired by the infant cognition literature (Shu et al., 2021; Smith et al., 2019), and also contained 3 hand-animated scenarios. In total, there were 28 scenarios, where each scenario consisted of 3 videos: a familiarization video, an unexpected outcome video, and an expected outcome video. We assigned these videos to the physical domain or psychological domain. The 16 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**) (Figure 1A). The 20 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* (Figure 1B), or surprising *environments* (Figure 1C) in which the actions 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**), or moved as though they were circumventing an obstacle, which was then missing (**infer-constraint**).

Experimental design 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. Each trial had 4 parts: a familiarization movie (7.5s), a corresponding test movie (7.5s; either unexpected or expected), a fixation cross for a jittered duration of 4-10s, and an attention check (2s). Participants were asked to attend to the movies. During the attention check, they saw an image of an agent, object, or surface texture, and responded via button press whether that image appeared in the most recent trial. In anticipation that we may need to restrict our analysis to the first 2 runs (see Results), scenarios were split into two halves, one half assigned to runs 1-2 and the other assigned to runs 3-4. We generated 128 unique random event sequences, one per run per subject, 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.

Data acquisition, pre-processing and analysis

Data were acquired from a 3-Tesla Siemens Magnetom Prisma scanner located at the Athinoula A. Martinos Imaging Center at MIT, using a 32-channel head coil. Participants viewed stimuli through a mirror projected to a 12” x 16” screen behind the scanner, at a visual angle of approximately 14 x 19 degrees. Participants underwent 2 runs of each localizer task, and 4 runs of the VOE task, presented in an interleaved and fixed order. Neuroimaging data were pre-processed using fmriprep (Esteban et al., 2019), and then analyzed using in-house lab scripts, described in detail at `https://osf.io/sa7jy/`. Please see our pre-registration for details about data acquisition parameters, pre-processing, and first-, second-, and group-level analysis.

Subject-specific functional regions of interest (ssfROI) analysis The goal of our subject-specific functional regions of interest (ssfROI) analysis (Fedorenko, Hsieh, Nieto-Castañón, Whitfield-Gabrieli, & Kanwisher, 2010) is to find, in individual participants, voxels that are maximally engaged during processing of low-level visual features and visual motion (motionLoc), goal-directed attention (spWMloc), and social and physical processing (DOTSloc) while allowing the precise voxels selected to vary somewhat across people, to fit their unique neuroanatomy and function. For each localizer task, for each participant, we selected the top 100 voxels (i.e., those with the highest z values for the corresponding contrasts listed in Figure 2B) from each participant's second-level map, within functional search spaces (Figure 2A) derived from previous work (multiple demand parcels from `https://evlab.mit.edu/funcloc/`; V1 and MT parcels from Pramod et al., 2022; and domain-specific parcels from our prior experiment).

Then, given these regions of interest (ROIs) for each subject, we measured the response of these voxels to all conditions from our primary VOE task (i.e. that condition > rest) for each of the 4 runs. For our confirmatory analysis, we focused on the unexpected and expected test events from physical scenarios, and psychological scenarios involving surprising actions (Figure 1A-B). In each region, we modeled responses predicted by domain (psychology vs physics), event (unexpected vs expected), and, for domain-specific regions, an interaction between them. In further exploratory analyses, we studied the responses of these regions to psychological violations that involved surprising environments (psychology-environment scenarios; Figure 1C).

Results

Following our analysis plan, we conducted a manipulation check, asking whether the VOE effect (unexpected > expected) declined across runs. Indeed, folding across all regions for which we predicted a VOE effect, the average VOE effect was more evident in runs 1 ($p = 0.07$) and 2 ($p = 0.04$) than in runs 3 ($p = 0.74$) or 4 ($p = 0.69$). Thus we followed our plan to restrict all subsequent confirmatory analyses to the first two runs. We used linear mixed effects models implemented in *lme4* (Bates, Mächler, Bolker, & Walker, 2015), and modeled the average response per region as predicted by a main effect of domain, a main effect of event, and (for domain-specific regions) the interaction across them. We used the *lsmeans* package (Lenth, 2016) to return pairwise estimates from interactions. Model formula: $meanbeta \sim$

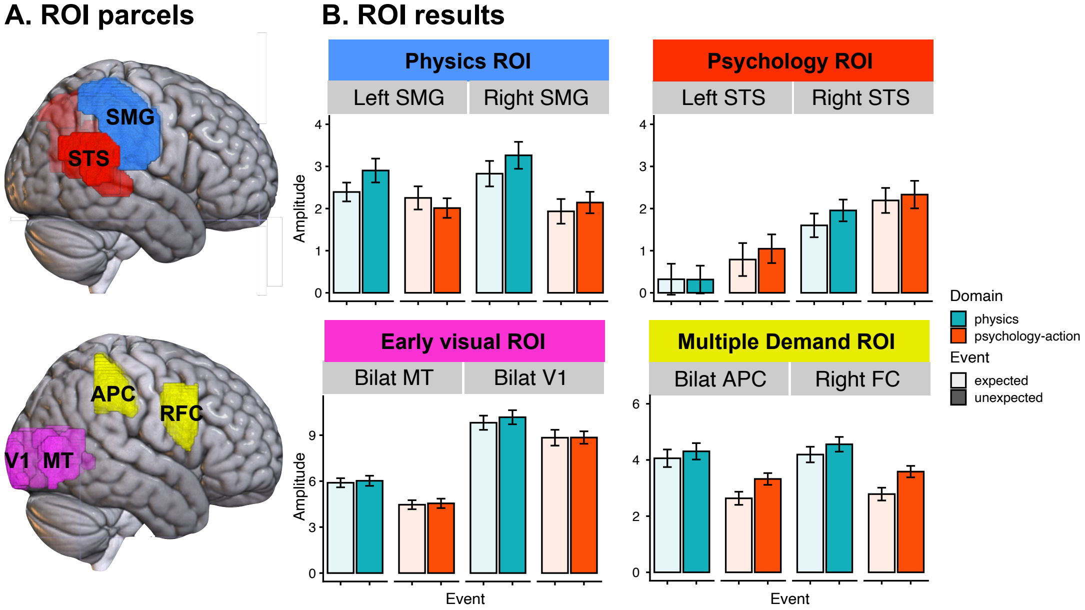

Figure 2: Results of subject-specific functional regions of interest (ssfROI) analysis. (A) Parcels for all focal regions of interest. (B) ssfROI results in domain-general multiple demand and early visual regions (bottom row; bilateral anterior parietal cortex (APC), right frontal cortex (RFC), bilateral primary visual cortex (V1) and bilateral motion sensitive area (MT)), and domain-specific regions (top row: left and right superior temporal sulcus (STS), left and right supramaringal gyrus (SMG)). Y axis indicates the average beta (i.e. magnitude of response) per region, relative to fixation, across *N*=32 participants. Error bars indicate standard error of the mean, taking into account within-subjects variance.

$domain * event + (1|subjectID)$. Full regression tables for all analyses are available at `https://osf.io/sa7jy/`. Our significance threshold for these analyses was p < .025, two-tailed, corrected for the number of regions (2) per hypothesis.

Confirmatory results: Physics and Psychology-Action events

Our **early visual** regions of interest were bilateral primary visual area (V1) and bilateral motion sensitive area (MT). Both regions responded more to physical than psychological events (V1: unstandardized coefficient (B) = 0.58 95% confidence interval (CI) [0.31, 0.84], p < .001; MT: B = 0.73 [0.59, 0.86], p < .001), but did not respond differently to unexpected and expected versions of these events (V1: B = -0.09 [-0.36, 0.17], p = .491; MT: B = -0.05 [-0.19, 0.08], p = 0.427).

Our **multiple demand** regions of interest were bilateral anterior parietal cortex (APC) and right frontal cortex (RFC). Both regions responded more responded more to physical than psychological events (APC: B = 0.60 [.40, 0.80], p < .001; RFC: B = 0.60 [0.41, 0.78], p < .001), and responded more to unexpected events, regardless of domain (APC: B = -0.23 [-0.44, -0.03], p = .024; RFC: B = -0.29 [-0.48, -0.10], p = .003)[1].

Our **domain-specific physics** regions were left and right supramarginal gyrus (SMG). Both left and right SMG responded more to physical than psychological events (LSMG: B = 0.26 [0.10, 0.41], p = .001; RSMG: B = 0.50 [0.33, 0.67], p < .001). Right SMG showed a marginally higher response to unexpected events regardless of domain, which did not survive correction for multiple comparisons (B = -0.16 [-0.33, 0.01], p = .061). Left SMG showed a signature of domain-specific prediction error: an interaction between domain and event (B=-0.19, [-0.34, -0.03], p = .018), with greater responses for unexpected than expected physical events (p = 0.023), and no significant VOE effect for psychological events (p = 0.280).

Our **domain-specific psychology** regions were left and right superior temporal sulcus (STS). Both left and right STS responded more to psychological events (LSTS: B = -0.30 [-0.49, -0.11], p = .002; RSTS: B = -0.24 [-0.41, -0.08], p = .004). However, neither left nor right STS responded more to unexpected than expected events (LSTS: B = -0.06 [-0.25, 0.13], p = .524; RSTS: B = -0.12 [-0.29, 0.04], p = .139), and there was no interaction between domain and event in these regions (LSTS: B = 0.07 [-0.13, 0.26], p = .501; RSTS: B = -0.05 [-0.22, 0.11], p = .517).

Exploratory results: Psychology-Environment Events

We then modeled the responses of all the above regions, to the psychological scenarios involving surprising environ-

[1] The selection process of these multiple demand ROIs was pre-registered, but due to a mistake in the analysis, the two regions that we originally selected, bilateral inferior frontal gyrus (IFG) and bilateral insula, are not the ones we report here. This deviates from our original pre-registration document, which we accordingly updated at `https://osf.io/8ywah`.

ments; Fig 1C), as predicted by event only (model formula: $meanbeta \sim event + (1|subjectID)$. The only region that responded more to unexpected than expected events in this category, in the first two runs (the same subset of the data as the confirmatory results) was the right STS (B = -0.319 [-0.54, -0.09], p = .006).

Discussion

A ball passing through a wall is visually unfamiliar, violates expectations about the physical world, and evokes curiosity about what caused that event to occur. Which of these processes likely explain adults' judgements of, and perhaps infants' longer looking to, these events? Here, we tested these hypotheses by scanning the brains of adults using fMRI, and studied the responses in regions that correspond to each hypothesis under investigation.

Early visual regions, V1 and MT, did not respond more to unexpected than expected VOE events, in either domain. This result suggests that the VOE response, in adults, is not explained by low-level visual prediction error that is propagated up to higher-level cortical regions. This result provides evidence against the claim that such events attract attention merely because they contain confounding low-level visual features (Sirois & Jackson, 2012; Rivera et al., 1999; Heyes, 2014).

Multiple demand regions, APC and RFC, responded more to unexpected events from both domains: The voxels from this region that were maximally engaged when each individual participant performed a difficult versus easy working memory task, also responded more when they viewed surprising than expected psychological and physical events. This result suggests that processes similar to goal-driven attention are support the VOE response in adults.

Lastly, one region implicated in physical processing, the left supramarginal gyrus (LSMG), responded to unexpected physical events, more than to expected physical events or any psychological events. Thus, in adults, the VOE responses for physical events is supported in part by mental processes that are dedicated to physical reasoning in particular.

In summary, of the hypotheses under consideration, we found evidence that VOE stimuli from developmental psychology evoked both domain-general and domain-specific processes: (i) domain-general goal-driven attention, but not early visual processing and (ii) for physical events, domain-specific, distinctively physical, processing.

Implications for studies of infants

Our approach——studying the brains of adults in order to evaluate hypotheses about neural function and behavior in infants——has both strengths and limitations. Because of continuity in the the organization of large-scale cortical networks (Eyre et al., 2021) between infants and adults, as well as the cortical responses evoked by agents and objects (Lloyd-Fox et al., 2009; Wilcox et al., 2010) and domain-general prediction error (Werchan et al., 2016; Nakano et al., 2009), we believe that some of these findings have implications for infant cognition. However, infants are not adults, and thus any conclusions we draw about infants, given neuroimaging data from adults, are subject to some assumptions about continuity that may or may not be justified (Blumberg & Adolph, 2023; Liu, Raz, Kamps, Grossmann, & Saxe, 2023).

The current ssfROI approach is a strength of our design, but difficult to implement in infants. With independent localizer tasks, we were on firmer ground to study two separate mental processes (goal-directed attention on the one hand; physical perception on the other) in the same cortical territory (e.g. SMG and APC). Adults can tolerate longer scans and can be instructed to perform tasks in the scanner, and it is much harder to design and run localizer tasks in infants. But without localizers, reverse inference over functional activation alone is not straightforward. Thus, while it may be possible to study neural responses in infants using the insights from this research, significant challenges remain.

Neural origins of rational action understanding

The superior temporal sulcus (STS) showed a surprising pattern of results. In prior literature, the STS, especially the right STS, responds to many aspects of agents, including voices, facial expressions, direction of gaze, action intentions and outcomes, and social interactions (Deen et al., 2015; Gao, Scholl, & McCarthy, 2012; Isik et al., 2017; Saxe et al., 2004; Walbrin, Downing, & Koldewyn, 2018; Shultz, Lee, Pelphrey, & McCarthy, 2011; Pelphrey, Viola, & McCarthy, 2004). In our study, we defined this region in individual participants by selecting the voxels, within a parcel in the posterior portion of STS (see Figure 2A), that maximally responded to social over physical interaction in an independent task (DOTSloc; Fischer et al., 2016). This region did not respond more to unexpected psychological events containing an unexpected action (goal, efficiency; see Marsh, Mullett, Ropar, & Hamilton, 2014; Ramsey & Hamilton, 2010 for similar findings from prior work). However, in our exploratory analyses, right STS did respond more to unexpected physical outcomes generated by an agent. This surprising pattern merits further investigation.

More broadly, we speculate that the computations that support understanding of rational action rely on input from other domains of reasoning. For example, representing the efficiency of an action may require physical computations (Liu, 2022), e.g. representing the agent and obstacle as solid bodies, and the agent's action as resisting gravity. Previous studies show that frontal regions implicated in action perception, and also parietal regions implicated in physical reasoning, both respond to deviations from goal-directed and efficient action (Marsh et al., 2014; Ramsey & Hamilton, 2010), including in infants (Southgate, Begus, Lloyd-Fox, di Gangi, & Hamilton, 2014). Although the focus of this research was on VOE, we look forward to using these data to study the interplay between intuitive physics and psychology.

Acknowledgments

We gratefully acknowledge the following funding sources: DARPA Machine Common Sense Program (CW3013552), NIH F32HD103363 (to SL).

References

Aslin, R. N. (2007). What's in a look? *Dev. Sci.*, *10*(1), 48–53.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Softw.*, *67*(1).

Blumberg, M. S., & Adolph, K. E. (2023). Protracted development of motor cortex constrains rich interpretations of infant cognition. *Trends Cogn. Sci.*.

Carey, S. (2009). *The origin of concepts*. Oxford University Press.

Deen, B., Koldewyn, K., Kanwisher, N., & Saxe, R. (2015). Functional organization of social perception and cognition in the superior temporal sulcus. *Cereb. Cortex*, *25*(11), 4596–4609.

Dehaene-Lambertz, G., & Spelke, E. S. (2015). The infancy of the human brain. *Neuron*, *88*(1), 93–109.

DiNicola, L. M., Braga, R. M., & Buckner, R. L. (2020). Parallel distributed networks dissociate episodic and social functions within the individual. *J. Neurophysiol.*, *123*(3), 1144–1179.

Emberson, L. L., Boldin, A. M., Robertson, C. E., Cannon, G., & Aslin, R. N. (2018). Expectation affects neural repetition suppression in infancy. *Dev. Cogn. Neurosci.*, 100597.

Esteban, O., Markiewicz, C. J., Blair, R. W., Moodie, C. A., Isik, A. I., Erramuzpe, A., ... Gorgolewski, K. J. (2019). fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nat. Methods*, *16*(1), 111–116.

Eyre, M., Fitzgibbon, S. P., Ciarrusta, J., Cordero-Grande, L., Price, A. N., Poppe, T., ... Edwards, A. D. (2021). The developing human connectome project: typical and disrupted perinatal functional connectivity. *Brain*, *144*(7), 2199–2213.

Farroni, T., Chiarelli, A. M., Lloyd-Fox, S., Massaccesi, S., Merla, A., Di Gangi, V., ... Johnson, M. H. (2013). Infant cortex responds to other humans from shortly after birth. *Sci. Rep.*, *3*, 2851.

Fedorenko, E., Duncan, J., & Kanwisher, N. (2013). Broad domain generality in focal regions of frontal and parietal cortex. *Proc. Natl. Acad. Sci. U. S. A.*, *110*(41), 16616–16621.

Fedorenko, E., Hsieh, P.-J., Nieto-Castañón, A., Whitfield-Gabrieli, S., & Kanwisher, N. (2010). New method for fMRI investigations of language: defining ROIs functionally in individual subjects. *J. Neurophysiol.*, *104*(2), 1177–1194.

Fischer, J., Mikhael, J. G., Tenenbaum, J. B., & Kanwisher, N. (2016). Functional neuroanatomy of intuitive physical inference. *Proc. Natl. Acad. Sci. U. S. A.*, *113*(34), E5072–81.

Frantz, R. L., Ordy, J. M., & Udelf, M. S. (1962). Maturation of pattern vision in infants during the first six months. *J. Comp. Physiol. Psychol.*, *55*(6), 907–917.

Gao, T., Scholl, B. J., & McCarthy, G. (2012). Dissociating the detection of intentionality from animacy in the right posterior superior temporal sulcus. *J. Neurosci.*, *32*(41), 14276–14280.

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

Gergely, G., Nádasdy, Z., Csibra, G., & Bíró, S. (1995). Taking the intentional stance at 12 months of age. *Cognition*, *56*(2), 165–193.

Henson, R., Shallice, T., & Dolan, R. (2000). Neuroimaging evidence for dissociable forms of repetition priming. *Science*, *287*(5456), 1269–1272.

Heyes, C. (2014). False belief in infancy: a fresh look. *Dev. Sci.*, *17*(5), 1–13.

Isik, L., Koldewyn, K., Beeler, D., & Kanwisher, N. (2017). Perceiving social interactions in the posterior superior temporal sulcus. *Proc. Natl. Acad. Sci. U. S. A.*, *114*(43), E9145–E9152.

Jiang, J., Summerfield, C., & Egner, T. (2016). Visual prediction error spreads across object features in human visual cortex. *J. Neurosci.*, *36*(50), 12746–12763.

Kang, M. J., Hsu, M., Krajbich, I. M., Loewenstein, G., McClure, S. M., Wang, J. T.-Y., & Camerer, C. F. (2009). The wick in the candle of learning: epistemic curiosity activates reward circuitry and enhances memory. *Psychol. Sci.*, *20*(8), 963–973.

Lenth, R. V. (2016). Least-squares means: The R package lsmeans. *Journal of Statistical Software*, *69*(1), 1–33. doi: 10.18637/jss.v069.i01

Liu, S. (2022). From infancy, we model people as minds, acting in a physical world. *Preprint at https://psyarxiv.com/sg4bz*.

Liu, S., Raz, G., Kamps, F., Grossmann, T., & Saxe, R. (2023). No evidence for discontinuity between infants and adults. *Preprint at https://psyarxiv.com/9yt5u; in press at Trends in Cognitive Sciences*.

Lloyd-Fox, S., Blasi, A., Volein, A., Everdell, N., Elwell, C. E., & Johnson, M. H. (2009). Social perception in infancy: a near infrared spectroscopy study. *Child Dev.*, *80*(4), 986–999.

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

Nakano, T., Watanabe, H., Homae, F., & Taga, G. (2009). Prefrontal cortical involvement in young infants' analysis of novelty. *Cerebral Cortex*, *19*(2), 455–463.

Needham, A., & Baillargeon, R. (1993). Intuitions about support in 4.5-month-old infants. *Cognition*, *47*(2), 121–

148.

Parris, B. A., Kuhn, G., Mizon, G. A., Benattayallah, A., & Hodgson, T. L. (2009). Imaging the impossible: an fMRI study of impossible causal relationships in magic tricks. *Neuroimage*, *45*(3), 1033–1039.

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

Peeles, D. R., & Teller, D. Y. (1975). Color vision and brightness discrimination in two-month-old human infants. *Science*, *189*(4208), 1102–1103.

Pelphrey, K. A., Viola, R. J., & McCarthy, G. (2004). When strangers pass: processing of mutual and averted social gaze in the superior temporal sulcus. *Psychol. Sci.*, *15*(9), 598–603.

Perez, J., & Feigenson, L. (2022). Violations of expectation trigger infants to search for explanations. *Cognition*, *218*, 104942.

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

Ramsey, R., & Hamilton, A. F. d. C. (2010). Triangles have goals too: understanding action representation in left aIPS. *Neuropsychologia*, *48*(9), 2773–2776.

Rivera, S. M., Wakeley, A., & Langer, J. (1999). The drawbridge phenomenon: representational reasoning or perceptual preference? *Dev. Psychol.*, *35*(2), 427–435.

Robertson, C. E., Thomas, C., Kravitz, D. J., Wallace, G. L., Baron-Cohen, S., Martin, A., & Baker, C. I. (2014). Global motion perception deficits in autism are reflected as early as primary visual cortex. *Brain*, *137*(Pt 9), 2588–2599.

Saxe, R., Xiao, D.-K., Kovacs, G., Perrett, D. I., & Kanwisher, N. (2004). A region of right posterior superior temporal sulcus responds to observed intentional actions. *Neuropsychologia*, *42*(11), 1435–1446.

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

Shu, T., Bhandwaldar, A., Gan, C., Smith, K. A., Liu, S., Gutfreund, D., … Ullman, T. D. (2021). AGENT: A benchmark for core psychological reasoning. *38th International Conference on Machine Learning (ICML), 2021*.

Shultz, S., Lee, S. M., Pelphrey, K., & McCarthy, G. (2011). The posterior superior temporal sulcus is sensitive to the outcome of human and non-human goal-directed actions. *Soc. Cogn. Affect. Neurosci.*, *6*(5), 602–611.

Sim, Z. L., & Xu, F. (2017). Infants preferentially approach and explore the unexpected. *Br. J. Dev. Psychol.*.

Sim, Z. L., & Xu, F. (2018). Another look at looking time: Surprise as rational statistical inference. *Top. Cogn. Sci.*.

Sirois, S., & Jackson, I. R. (2012). Pupil dilation and object permanence in infants. *Infancy*, *17*(1), 61–78.

Smith, K., Mei, L., Yao, S., Wu, J., Spelke, E., Tenenbaum, J., & Ullman, T. (2019). Modeling expectation violation in intuitive physics with coarse probabilistic object representations. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems 32* (pp. 8983–8993). Curran Associates, Inc.

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

Spelke, E. S. (2022). *What babies know: Core knowledge and composition volume 1*. Oxford University Press.

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

Stahl, A. E., & Feigenson, L. (2015). Cognitive development. observing the unexpected enhances infants' learning and exploration. *Science*, *348*(6230), 91–94.

Stahl, A. E., & Feigenson, L. (2019). Violations of core knowledge shape early learning. *Top. Cogn. Sci.*, *11*(1), 136–153.

Stahl, A. E., & Kibbe, M. M. (2022). Great expectations: The construct validity of the violation-of-expectation method for studying infant cognition. *Infant Child Dev.*.

van Lieshout, L. L. F., Vandenbroucke, A. R. E., Müller, N. C. J., Cools, R., & de Lange, F. P. (2018). Induction and relief of curiosity elicit parietal and frontal activity. *J. Neurosci.*, *38*(10), 2579–2588.

Walbrin, J., Downing, P., & Koldewyn, K. (2018). Neural responses to visually observed social interactions. *Neuropsychologia*, *112*, 31–39.

Werchan, D. M., Collins, A. G. E., Frank, M. J., & Amso, D. (2016). Role of prefrontal cortex in learning and generalizing hierarchical rules in 8-Month-Old infants. *J. Neurosci.*, *36*(40), 10314–10322.

Wilcox, T., Haslup, J. A., & Boas, D. A. (2010). Dissociation of processing of featural and spatiotemporal information in the infant cortex. *Neuroimage*, *53*(4), 1256–1263.

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