Functional neuroanatomy of intuitive physical inference

Jason Fischer[a,b,c,d,1], John G. Mikhael[a,b,c,e], Joshua B. Tenenbaum[a,b,c], and Nancy Kanwisher[a,b,c,1]

[a]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139; [b]McGovern Institute for Brain Research, Massachusetts Institute of Technology, Cambridge, MA 02139; [c]The Center for Brains, Minds, and Machines, Massachusetts Institute of Technology, Cambridge, MA 02139; [d]Department of Psychological and Brain Sciences, Johns Hopkins University, Baltimore, MD 21202; and [e]Harvard Medical School, Boston, MA 02115

Contributed by Nancy Kanwisher, June 29, 2016 (sent for review May 24, 2016; reviewed by Susan J. Hespos and Doris Tsao)

To engage with the world—to understand the scene in front of us, plan actions, and predict what will happen next—we must have an intuitive grasp of the world's physical structure and dynamics. How do the objects in front of us rest on and support each other, how much force would be required to move them, and how will they behave when they fall, roll, or collide? Despite the centrality of physical inferences in daily life, little is known about the brain mechanisms recruited to interpret the physical structure of a scene and predict how physical events will unfold. Here, in a series of fMRI experiments, we identified a set of cortical regions that are selectively engaged when people watch and predict the unfolding of physical events—a "physics engine" in the brain. These brain regions are selective to physical inferences relative to nonphysical but otherwise highly similar scenes and tasks. However, these regions are not exclusively engaged in physical inferences per se or, indeed, even in scene understanding; they overlap with the domain-general "multiple demand" system, especially the parts of that system involved in action planning and tool use, pointing to a close relationship between the cognitive and neural mechanisms involved in parsing the physical content of a scene and preparing an appropriate action.

physical scene understanding | mental simulation | fMRI | premotor cortex | action planning

Understanding, predicting, and acting on the world requires an intuitive grasp of physics (Fig. 1). We see not just a table and a coffee cup, but a table supporting a coffee cup. We see not just a ping pong ball moving after contact with a paddle, but the paddle causing the ball to move by exerting a force through that contact. We use physical intuitions to not just understand the world but predict what will happen next—that a stack of dishes is unstable and likely to fall or that a squash ball is on a trajectory to ricochet off the wall and head in our direction. We also need rich physical knowledge to plan our own actions. Before we pick up an object, we must assess its material and weight and prepare our muscles accordingly. To navigate our environment, we need to determine which surfaces will support us (e.g., a linoleum floor but maybe not the surface of a frozen stream; this tree branch but probably not that one) and what barriers are penetrable (e.g., a beaded curtain but not a glass wall). How do we compute these everyday physical inferences with such apparent ease and speed?

Battaglia et al. (1) recently proposed a computational mechanism for how humans can make a wide range of physical inferences in natural scenes via a mental simulation engine akin to the "physics engines" used in many video games. Physics engines are software systems that support efficient but approximate simulations of rigid body, soft body, or fluid mechanics for the purpose of generating realistic interactive gameplay in a virtual physical world. Rather than striving for fine-grained physical accuracy, game physics engines make shortcuts to capture dynamic interactions that "look good" to people over a wide range of situations and that can be generated in real time, often exploiting specialized hardware acceleration [such as graphics processing units (GPUs)]. Here, we ask: does the human brain possess an analogous "intuitive physics engine"—a brain region or family of regions essentially engaged in physical inferences and recruited more for physical inference than for other similarly difficult prediction or perception tasks?

Although some studies have explored the neural representation of objects' surface and material properties (2–4) and weights (5–7) or investigated the brain areas involved in explicit, textbook-style physical reasoning (8, 9), little is known about the cortical machinery that supports the more implicit perceptual judgments about physical events that are so pervasive in daily life. However, behavioral findings from young children and adults suggest that we use systematic cognitive and neural machinery to make physical inferences. During the first year of life, infants acquire a rich array of physical knowledge in a consistent order; 3- to 4-mo-old infants understand that the world is composed of bounded, unitary objects (10) that are continuous in time and space (11). By 5 mo of age, most infants are able to differentiate a liquid from a solid using motion cues and have expectations about how nonsolids behave and interact (12, 13); by 6 or 7 mo, they are sensitive to the causal roles of one object striking and launching another (14, 15), and by 8 mo, they can determine which objects must be attached for a configuration to be stable under gravity (16). By 12 mo, they are sensitive to the rough location of an object's center of mass, relative to the edge of a supporting surface, needed for that support relationship to be stable (17). These findings and other behavioral findings from young children suggest that humans may possess, from a young age, a mental framework for interpreting and learning about physical events (18, 19).

Although infant research has long emphasized the development of basic competences entailed in implicit physical understanding of simple scenes, early research on physical intuitions in adults

Significance

Perceiving the physical structure of the world and predicting how physical events will unfold over time are central to our daily lives. Recent behavioral and computational research has suggested that our physical intuitions may be supported by a "physics engine" in the brain akin to the physical simulation engines built into video games. However, to date, there has been almost no investigation of the brain areas involved in intuitive physical inference. Here, using fMRI, we show that a variety of physical inference tasks as well as simply viewing physically rich scenes engage a common brain network in frontal and parietal cortices. These findings open the door to the cognitive neuroscientific study of physical inference in the human brain.

Author contributions: J.F., J.B.T., and N.K. designed research; J.F. and J.G.M. performed research; J.F. analyzed data; and J.F., J.B.T., and N.K. wrote the paper.

Reviewers: S.J.H., Northwestern University; and D.T., California Institute of Technology.

The authors declare no conflict of interest.

[1]To whom correspondence may be addressed. Email: jason.fischer@jhu.edu or ngk@mit.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1610344113/-/DCSupplemental.

Fig. 1. Our experience of the world is shaped by our physical intuitions. (*A*) When carrying out an everyday task, like shopping for produce, we "see" which items can be safely removed from a pile without causing others to fall and which cannot. (*B*) When we encounter a heavy door, we know that pushing at the edge will be more effective in turning the door than pushing near the center. We are also aware of the family of possible physical outcomes of a scenario, and we form expectations about how likely different classes of outcomes are; for example, the laptop's position in *C* appears perilous, because there is a good chance that the laptop will eventually end up in the water or on the ground. However, the laptop's position in *D* causes little concern, despite the fact that it also rests near a large volume of water. We form these intuitions with apparent ease, and they constrain and interact with our goals (like choosing the perfect piece of fruit or keeping the laptop unharmed) to determine the actions that we take next.

used explicit, deliberate reasoning tasks and emphasized people's shortcomings relative to scientific norms. For example, when asked to diagram the path that a moving object will follow, people display systematic fallacies, sometimes drawing a curvilinear path for an object moving in the absence of external forces if it was previously moving in a rotational fashion (20) or drawing a straight path for an object falling toward the ground, even if it had forward momentum before being dropped (21, 22). More recent work, however, has focused on more implicit perceptual or action-based physical inference tasks of the kind that are critical in daily life, and these studies reveal that people can make sophisticated physical predictions if their intuitions are tested in this way. For example, misjudgments of a dropped object's path disappear if people act to catch the object rather than simply drawing its path (23). Indeed, many human judgments in perceptual tasks have been found to be quantitatively consistent with approximate probabilistic versions of Newtonian principles of motion (1, 24, 25). What neural resources do humans use to perform these inferences?

Here, we asked whether the brain has a physics engine—a brain region or set of regions, consistent across individuals, that implements fast intuitive physical inference from visually presented scenes. Three broad possible findings merit consideration. First, it is possible that no brain region is preferentially engaged in making physical inferences compared with other object or scene perception tasks matched on difficulty and visual content. Not all cognitive functions engage brain regions specialized for that function; indeed, many tasks rely on domain-general brain resources (26, 27), and physical inference may fall in this category. Second, physical scene understanding could rely on a domain-specific system specialized for this function in particular, analogous to specialized cortical systems identified for seeing faces (28), hearing speech (29), or thinking about other people's thoughts (30). Third, physical inference may engage brain regions also known to be involved in other functions, such as high-level vision (given the rich physical information included in visual scene understanding) or motor control (given the necessity of physical information for action planning). This work will help determine the anatomical consistency across individuals and functional specificity of the neural machinery underlying physical inference and may provide anatomical targets for future investigations of the dimensions of physical information encoded and the computations used to efficiently parse the physical content of a scene.

Results

Experiment 1: Physical and Nonphysical Judgments with Visually Identical Stimuli. In experiment 1, we screened broadly for candidate brain regions engaged in physical scene understanding by asking which, if any, regions responded more when participants judged the physical content of a stimulus than when they judged other visual content of the same stimulus. We used a variant of the block towers task in the work by Battaglia et al. (1): during scanning, participants viewed videos of unstable block towers and judged either where the blocks would land if the tower tumbled (physical judgment) or whether the tower contained more blue or yellow blocks (color judgment) (Fig. 2*A*). Critically, the stimuli presented for the two tasks were visually identical, and the tasks were matched on difficulty (*Materials and Methods*). Twelve participants each completed two runs of the task. We used one-half of the data from each subject (second run) to identify candidate functional regions of interest (fROIs) that showed a stronger response to the physics task than the color task. Specifically, we used the group-constrained subject-specific region of interest (ROI) definition method (31, 32). This approach identified regions ("parcels") of the cortex where many subjects had overlapping activations in the physics > color contrast, reflecting neighborhoods of common activation across subjects (Fig. 2*B*). For each individual subject, we then defined subject-specific fROIs by finding the subject's significant voxels within each of the group parcels. In this way, fROI locations were allowed to vary across individuals but required to fall within the same parcel to be labeled as a common ROI across subjects. This approach provided an objective and automatic means of localizing individual subject fROIs and establishing a common fROI labeling scheme across subjects without requiring voxelwise overlap in activations across subjects (additional details are in *Materials and Methods*). Subsequent analyses were performed within the fROIs defined individually within each subject. This approach yielded 11 distinct cortical parcels, most of which appeared in bilateral pairs. (Subcortical structures and the cerebellum were also included in the parcel generation process, but no consistent group activity appeared in those areas.) We labeled the parcels P1–P6 with an L or R hemisphere designation but analyzed all 11 parcels separately in subsequent analyses.

We validated and quantified the physics-related responses in the 11 parcels using the independent, left-out data from each subject's first run, which provides a statistically unbiased measure of the response magnitude of each region in each condition. Time courses for three example fROIs are shown in Fig. 2*C*, and response magnitudes (Fig. 2*D*) show a robust response for each fROI in the physics task but little response to the color task, despite the fact that the exact same stimuli were presented for the two conditions; only the task is driving the difference. Responses were significantly greater for the physics task than the color task in each fROI (Fig. 2). These 11 fROIs, thus, become candidates for brain regions engaged preferentially in physical inference, worthy of additional investigation.

Fig. 2*E* shows a random effects group activation map for the same contrast using all data (two runs from each subject). Significant voxels in the group random effects analysis generally fall

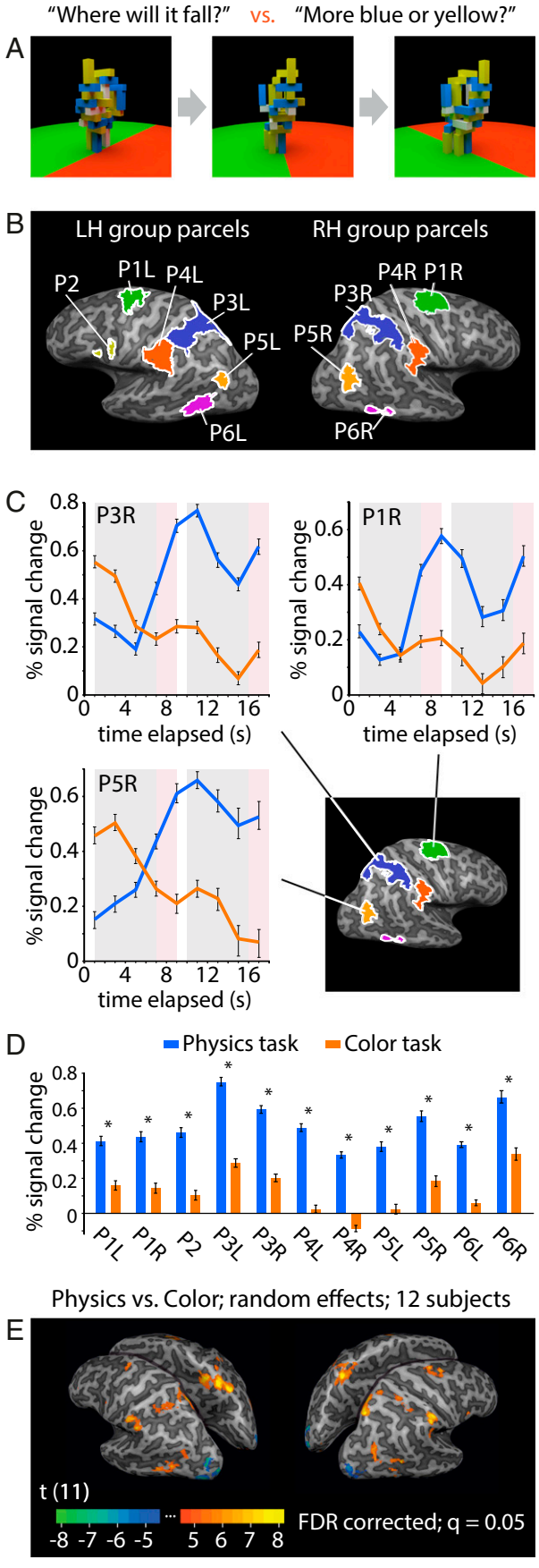

Fig. 2. Experiment 1 stimuli and results: physical vs. nonphysical judgments with visually identical stimuli. (*A*) During scanning, participants viewed movies of towers of blocks from the viewpoint of a camera panning around the tower. On each block, subjects were cued to report either (*i*) where more blocks would land (red vs. green side of the floor) if the tower tumbled or (*ii*) whether there were more blue or yellow blocks in the tower, ignoring white blocks. (*B*) Group parcels for the physics > color contrast computed using one run per subject. (*C*) Mean PSC in the blood-oxygen level-dependent response for three example parcels based on independent data from those used to define parcels and fROIs. Gray shading indicates the time during which the movies were shown, and pink shading indicates the response period during which a blank black screen was shown. Two videos with the same task instruction were shown per block; time courses show the full duration of one block. All fROIs showed a robust response during the physics task but little response to the color task. *D* shows mean PSC for all parcels. Response to the physics task was significantly greater than response to the color task in every parcel (t_{11} = 4.16, 4.50, 5.52, 7.84, 7.40, 8.55, 8.48, 5.58, 5.14, 8.33, and 3.75; P = 0.0016, 0.00091, 0.00018, 8×10^{-6}, 1.4×10^{-5}, 3×10^{-6}, 4×10^{-6}, 0.00017, 0.00032, 4×10^{-6}, and 0.0032 for P1L, P1R, P2, P3L, P3R, P4L, P4R, P5L, P5R, P6L, and P6R, respectively; paired t tests). Error bars in *C* and *D* are the bootstrapped SE across subjects. *Significant at q = 0.05 after false discovery rate correction for 11 comparisons. (*E*) Group random effects map for the physics > color contrast based on all data (two runs per subject).

within the parcels identified in the main (parcel-based) analysis but not always vice versa. Indeed, a major strength of the parcel-based method is that it can detect activations (e.g., the fusiform face area) that are present in nearby regions in most subjects but that often do not reach significance in a standard random effects analysis because of insufficient overlap across subjects (31). Another strength of the parcel-based method over the standard group analysis method is that the former leaves some data out for subsequent independent quantification of fROI responses, whereas the latter method standardly does not.

Given the use of identical stimuli and a task matched for difficulty, there was no a priori guarantee that any brain regions would respond more strongly in the physical inference task than in the color judgment task. However, experiment 1 revealed a systematic set of regions that responds at least twice as strongly when participants predict the physical outcome of a scenario than when they judge visual information (color) in the identical scenario. No difference in stimulus attributes could be responsible for the large effects evident in Fig. 2*D*, but it remains possible that differences in task content not related to physical reasoning per se drove the larger responses to the physical task. The physical task required mental simulation of what would happen next when the tower tumbled, whereas the color task did not; a general process of mental simulation (not specific to physical simulation) could be driving responses in the candidate regions. In this respect, experiment 1 is similar to a previous study that compared explicit physical reasoning with social reasoning in the brain (9). That study contrasted assessments of future physical outcomes with assessments of present mental states, and as in experiment 1 here, responses that they found could actually be because of prediction more generally. Importantly, however, experiment 1 and our subsequent tasks were designed to reflect the perceptual, action-based nature of physical inference in daily life (for which humans can be highly accurate), whereas the tasks used by Jack et al. (9) were explicit physical reasoning problems more akin to those studied by McCloskey et al. (20), for which observers show systematic errors. Similarly, another recent study by Mason and Just (8) presented textbook-based physics concepts as a way to investigate the neural representation of abstract information that must be learned through formal training. In contrast to these studies, the tasks that we use here are designed to engage the kind of intuitive physical inference that can be performed by preverbal infants. The physical task in experiment 1 also demanded attention to the spatial locations of particular blocks, because the exact arrangement determined how the blocks would fall. In the color task, the positions of individual blocks were irrelevant. Thus,

attention to spatial content could also be contributing to responses in the candidate regions. We, therefore, conducted a second experiment to control for these task differences.

Experiment 2: Physical vs. Social Interactions. In experiment 2, subjects watched pairs of dots moving within a square arena, with motion that implied either social interaction [like the classic animations by Heider and Simmel (33)] or physical interaction (bouncing like billiard balls) (Fig. 3*A*). In both cases, the dots moved within the arena for 8 s; then, one dot became invisible but remained within the arena, moving and interacting with the other dot for the final 2 s of the video. Participants were asked to predict the continuing trajectory of the now-invisible dot. On the final movie frame, when the dot reappeared, participants reported whether it had reappeared in an appropriate location. As in experiment 1, the judgments for the two movie types were matched on difficulty (*Materials and Methods*). Experiment 2 provided a complementary case to experiment 1: both the physical and social conditions used the same task, requiring mental simulation of spatial paths, but one implicitly invoked physical prediction and the other implicitly invoked social prediction.

Fig. 3*B* shows time courses of response for the physical and social tasks for the same fROIs shown in Fig. 2*C*. Example parcels P1R and P3R show a robust response timed to the onset of each of the two movies in the block for the physical motion and a substantially smaller and delayed response to the social interaction condition. By contrast, in P5R, the signal change was larger for the social condition than the physical condition. These effects are quantified in Fig. 3*C*, which shows the average response magnitude within each of the 11 parcels. Five parcels showed significantly greater response to the physical condition than the social condition: bilateral parcels in dorsal premotor cortex and supplementary motor area (P1L and P1R) and bilateral parcels in parietal cortex situated in somatosensory association cortex and the superior parietal lobule (P3L and P3R) as well as the left supramarginal gyrus (P4L). As in experiment 1, a random effects group contrast revealed significant voxels in the locations of these parcels (Fig. 3*D*) but may underestimate the extent of the cortex engaged by the task because of anatomical variability across subjects.

Importantly, the social condition in experiment 2 was not devoid of physical content; for example, the dots could not pass through walls or each other, and momentum was implied in the smooth motion. Even a region that was perfectly selective for physical inference (hence, not responsive to social content) would still be expected to respond somewhat to the social condition in experiment 2 because of this physical content. Indeed, we did see reliable responses to the social condition (compared with baseline) in all fROIs, especially during the simulation phase, but critically, responses to the physics condition were significantly greater than responses to the social condition in the five fROIs listed above, implicating them in the processing of physical information. The remaining fROIs showed robust responses to both conditions; these areas may perform general prediction or spatial functions that are necessary for the task in both conditions. Parcel P5R, which falls near cortical areas implicated in biological motion perception (34), showed a significantly greater response to the social condition than the physical condition. This finding shows that our design had sufficient power to uncover preferential responses to the social condition in expected regions.

The combined results of experiments 1 and 2 isolated five candidate regions that respond to physical content in both task and stimulus manipulations. The timing of the physics-related responses in experiment 2 also provides a clue to the components of the task that most effectively drive responses in these regions. The signal increased during the observation periods of the two physical videos in each block (gray shaded periods in Fig. 3*B*), when subjects were viewing the motion and collisions of the dots (allowing for standard hemodynamic lag). The signal did not increase appreciably after the physical prediction periods (shaded in pink in Fig. 3*B*), in which subjects actually had to mentally simulate the behavior of the hidden dot. Thus, it may be that these candidate regions can be engaged simply by the observation of physical content, even in the absence of the conscious, explicit effort to predict how physical events will unfold. Although the responses to the social condition were significantly smaller than those to the physical condition overall, the social condition did show an increase in response late in the observation period and during the simulation period. This delayed response may be because of participants using some physical constraints to predict the future behavior of the dots after inferring their social goals (e.g., observing that the red dot is chasing the blue dot and formulating a prediction that includes physical constraints, such as "the red dot will move in the direction of the blue dot, avoiding the barrier in the center of the arena").

If the observation of physical interactions drives responses in these candidate regions, this crucial role of observation might explain why our findings differ from those of a previous study that failed to find brain regions that were preferentially recruited for judging physical causality vs. social causality (35). Here, we used much longer observation periods, which allowed subjects to track physical behavior over the course of several seconds. To test the possibility that observation of physical events alone is sufficient to drive responses in the areas that we uncovered, we turned to an experiment in which subjects passively viewed movies that contained rich physical content.

Experiment 3: Passive Viewing of Physical Events. The results of experiment 2 suggest that simply observing physical events, such as objects colliding, falling, or rolling, may be sufficient to engage the brain's physics engine. To test whether passive viewing of physical events elicits responses in the candidate regions uncovered in experiments 1 and 2, we conducted a new analysis on a large dataset that had been collected previously [the results of which were published in the work by Julian et al. (31)]. The experiment was originally designed as a localizer for face-, object-, scene-, and body-selective cortical areas using passively viewed 3-s video clips (Fig. 4*A*). We posited that, if the perceived physical content of the movies differed across categories, these differences might be reflected in the degree to which viewing them engages the same regions identified in experiments 1 and 2 if passive viewing of physical stimulus content is sufficient to drive responses in those regions.

We first collected ratings of the physical content in the movies from 30 workers on Amazon Mechanical Turk (AMT) (*Materials and Methods*). Fig. 4*A* shows the physical content rating for each of five categories ordered from highest to lowest. Statistically reliable differences were found between categories ($F_{4,295} = 127.95$; $P = 3.37 \times 10^{-63}$; one-way ANOVA), with object movies showing the highest physical content ratings. Furthermore, object movies were rated significantly higher than scrambled objects ($t_{118} = 9.48$; $P = 3.4 \times 10^{-16}$; two-sample t test), showing that, although the scrambling procedure maintained the low-level visual content and motion of the original movies, it diminished the ability to perceive physical interactions within those movies. To test whether the physical content in the passively viewed object movies engaged brain regions similar to the physics-responsive areas identified in experiments 1 and 2, we, therefore, examined the contrast of objects > scrambled objects in the fMRI data. Note that this contrast is also a standard localizer for object- and shape-selective cortical regions and will reveal known areas that are selective for those properties. Our questions here were whether the passively viewed physical content in the dynamics of the object movies was sufficient to engage additional regions beyond the classic object-selective areas and whether any additional regions align with those found in experiments 1 and 2. Using one-half of the data from each subject (even runs), we identified group parcels and corresponding individual subject fROIs for the objects > scrambled objects contrast using the same group-constrained subject-specific procedure as in experiment 1.

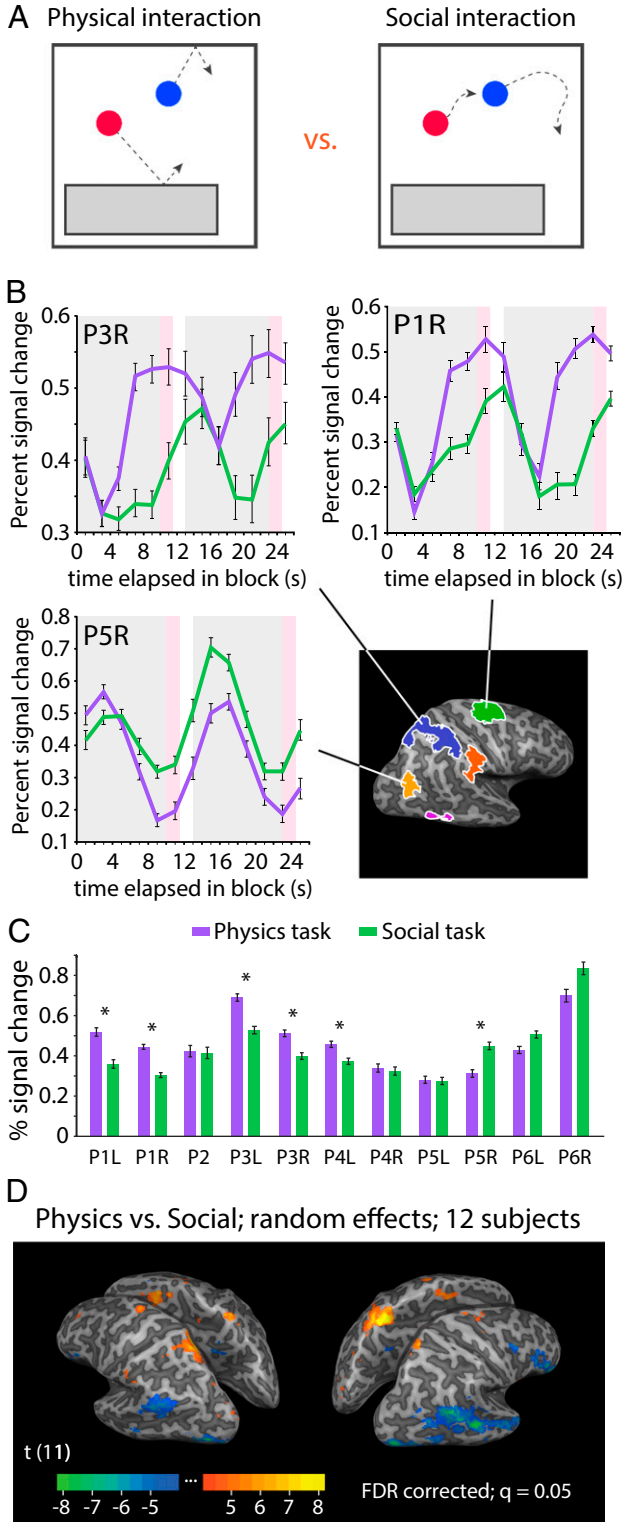

Fig. 3. Experiment 2 stimuli and results: physical vs. social interactions. (*A*) During scanning, participants viewed 10-s movies of dots moving around arenas. The motion of the dots indicated either physical interactions or social interactions, and in each case, the participant imagined where one of the dots would travel during a 2-s period when it was invisible. (*B*) Mean percentage change in the blood-oxygen level-dependent signal over the course of a block for three example parcels. Data were analyzed within the same individual subject ROIs defined in experiment 1. Two videos with the same task instruction were shown per block during seconds 1–10 and 14–23 of the 26-s block. (*C*) Only a subset of the parcels showed a stronger response during viewing and imagining physical interactions vs. social interactions (t_{11} = 3.67, 5.57, 0.17, 4.18, 2.98, 2.63, 0.31, 0.14, 3.48, 2.07, and 2.02; P = 0.0037, 0.0002, 0.87, 0.0015, 0.012, 0.023, 0.76, 0.89, 0.0052, 0.063, and 0.069 for P1L, P1R, P2, P3L, P3R, P4L, P4R, P5L, P5R, P6L, and P6R, respectively; paired t tests). *Significant at q = 0.05 after false discovery rate correction for 11 comparisons. (*D*) Group random effects map for the physical interactions > social interactions contrast. Note the significant social responses in expected areas along the superior temporal sulcus.

Fig. 4*B* shows the group parcels identified by the objects > scrambled objects contrast. In addition to parcels that fell in expected object-selective locations in visual cortex and along the ventral temporal surface, five additional parcels (highlighted in blue in Fig. 4*B*) appeared in locations that overlapped substantially with the candidate physics-responsive regions that we found in experiments 1 and 2. To characterize the response across all five stimulus categories within these parcels, we examined the signal change within the independent, left-out data (odd runs). The signal change in these parcels for the five stimulus categories corresponded closely to the independently collected ratings of physical content, falling in exactly the same order in three of five parcels. Thus, within the candidate physics-responsive parcels, the level of fMRI response to passively viewed movies is well-predicted by the richness of the physical content in those movies. In addition to the object movies, the body movies elicited strong positive responses across the candidate physics regions, which might be expected given that physical constraints factor critically into computational models for planning body movements (36), and similar networks have been suggested as the neural substrate of a physically integrated body schema (37).

These results show that the conscious, explicit effort to mentally simulate the physical behavior of the objects in a scene is not required to robustly engage these regions. However, these findings do not imply that physical scene processing necessarily happens automatically all of the time, irrespective of a person's goals or attentional state. Indeed, in experiment 1, selectively attending to the tower's color and ignoring its physical stability eliminated responses in the candidate physics-responsive regions. Although passive viewing of physical scenarios is sufficient to engage the brain's physics engine, a demanding competing task may be able to draw resources away, such that physical scene content is processed less deeply or automatically. The results of experiment 4 also do not imply that these brain areas do not play a key role in the mental simulation of physical outcomes, but they may do so in an online and automatic fashion, generating expectations that guide behavior as the events in a scene unfold.

Experiment 4: Relationship to the Multiple Demand Network and Motor Planning. The previous experiments used difficulty-matched tasks (or in the case of experiment 3, no task at all). The fact that the same set of physics-responsive regions emerged consistently in these difficulty-controlled experiments indicates that it is not just general mental effort driving the responses in these regions. Still, it could be the case that intuitive physical inference is carried out by strictly the same domain-general cortical regions that contribute to a wide variety of tasks, termed the multiple demand (MD) network (38). Responses in the MD areas generally scale with task difficulty, and this network is thought to provide a flexible problem-solving framework that contributes to general intelligence (38). To test whether the physics-responsive areas identified in the first three experiments are the same as the MD network, we separately localized the MD network in the same 12 subjects who participated in the first two experiments. During scanning, these subjects performed spatial working memory and verbal working memory tasks based on those in the work by Fedorenko et al. (27), which contrasted hard (high-load) vs. easy (low-load) conditions (task details are in Fig. S1). Fig. 5*A* shows the pattern of response for

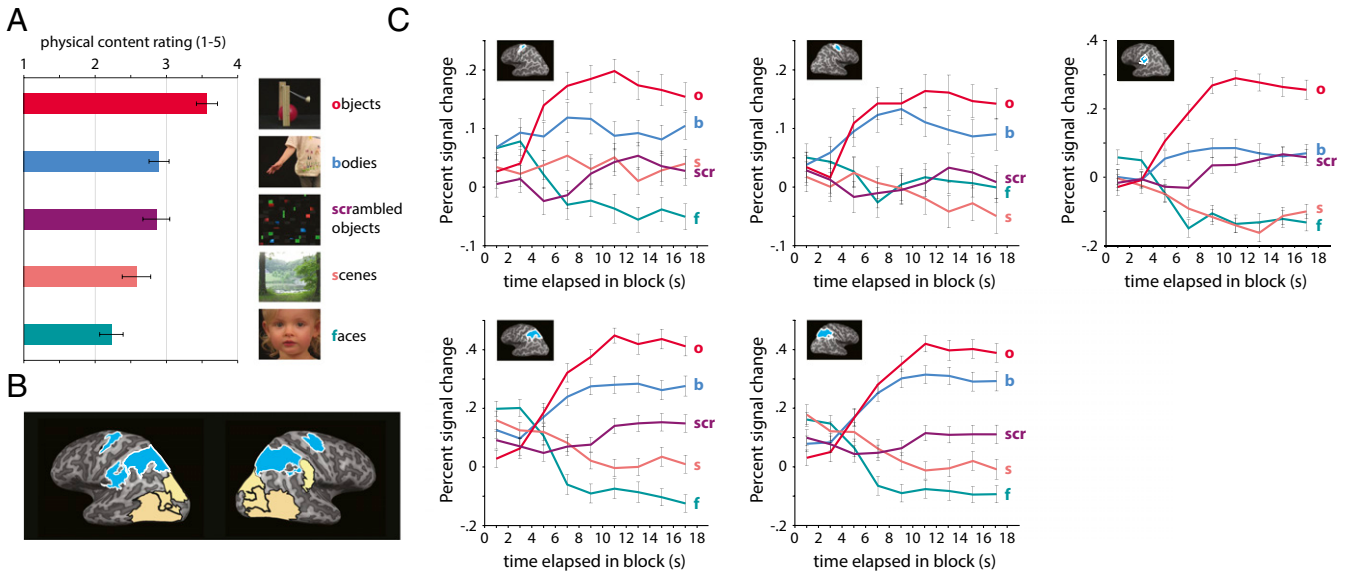

Fig. 4. Experiment 3 results: passive viewing of physical events. (*A*) We analyzed existing data from an experiment in which 65 participants passively viewed 3-s video clips containing objects, bodies, scrambled objects, scenes, and faces. We separately obtained ratings from 30 workers on AMT who rated the degree of physical content in the movies on a scale from one (least physical content) to five (most physical content) (*Materials and Methods*). Bars show the mean physical content rating for each of five categories; error bars are ±SE across subjects. (*B*) Group parcels generated based on an objects > scrambled objects contrast in one-half of the data from each subject (runs 2 and 4). Five parcels, highlighted in blue, overlapped substantially with physics-responsive parcels identified in experiments 1 and 2. (*C*) PSC plots for five highlighted parcels computed from data in independent runs (runs 1 and 3). The signal change in these parcels for five stimulus categories corresponded closely to the independently collected ratings of physical content, with objects receiving the highest physical content ratings and producing the largest signal change and faces receiving the lowest physical content ratings and producing the smallest signal change. Thus, within the candidate physics-responsive parcels, the level of blood-oxygen level-dependent response to passively viewed stimuli is well-predicted by the physical content in the stimuli.

each task in the hard > easy contrast. The two tasks showed a highly similar pattern of difficulty modulation across the brain, despite the differences in stimulus content, reflecting the previously described domain generality of the MD network (38). This pattern of response overlaps substantially with the areas that we have found to be engaged by physical inference (44.7 ± 6.0% of the voxels in the physical inference ROIs showed significant responses in both the spatial working memory and verbal working memory contrasts) but also, seems to include additional cortical areas that were not physics-responsive in our experiments. To test the similarity of MD network responses and intuitive physics-related responses (that is, whether they are likely to engage identical sets of areas, allowing for some noise), we computed the correlation between the whole-brain working memory and physical inference maps and compared the strength of this correlation with the correlation between the maps for the two working memory tasks and the correlation between the maps for the two physical inference tasks from experiments 1 and 2 (*Materials and Methods*). Fig. 5*B* shows these correlations: although we found a significant correlation between the spatial pattern of activation for the physics tasks and the working memory tasks ($t_{11} = 5.13$; $P = 0.00033$; one-sample t test), the activation pattern was more similar between the two working memory tasks ($t_{11} = 2.87$; $P = 0.015$; paired t test) and between the two physics tasks ($t_{11} = 3.04$; $P = 0.011$; paired t test). This pattern of results indicates that the physical inference-related activations and the MD activations are similar to each other but also, significantly different from each other.

More specifically, the physics-responsive regions seem to sit within a subset of the MD network. What distinguishes this subset from the rest of the MD network? Fig. 5*C* shows group parcels for the MD network generated based on the hard > easy contrasts from the spatial and verbal working memory tasks. The parcels are colored to reflect the magnitude of the towers > color contrast from experiment 1. The subset of the MD network most strongly engaged by physical inference resembles the brain regions discussed in the literatures on motor planning (39–43) and tool use (44–46) [figure 1 in the work by Gallivan and Culham (40) shows a meta-analysis]. This overlap points to the intriguing possibility of shared functional neuroanatomy for physical scene understanding and action planning. However, despite the apparent overlap among brain regions previously implicated in multiple demands, motor planning, and tool use, these literatures rarely engage with each other and use different experimental paradigms in distinct groups of subjects. As a consequence, it is difficult to determine whether these literatures refer to the same underlying neural system, and there is no straightforward way to determine with which system (if they are, indeed, distinct) the physical inference regions that we find are most closely associated. To do so will require running multiple paradigms from all three literatures in addition to physical inference paradigms within the same subjects, a substantial undertaking. At present, we simply note the striking overlap of physics-responsive regions and motor planning/tool use regions and the intriguing possibility that physical inference and motor function are intimately linked in the brain.

Discussion

This study found that physical scene understanding engages a systematic set of brain regions replicated across three studies: one holding the stimulus constant and varying the task (a tower-falling task vs. a color judgment task), one holding the task constant ("what will happen next?") and varying the stimuli (which had physical vs. social content), and one contrasting passive viewing of engaging movies that contained extensive physical content (e.g., colliding objects) vs. nonphysical content (e.g., faces). This systematic pattern of activation across all three tasks includes bilateral frontal regions (dorsal premotor cortex/supplementary motor area), bilateral parietal regions (somatosensory association cortex/superior parietal lobule), and the left supramarginal gyrus. This pattern of activation cannot be explained by generic task demands (because difficulty was matched across conditions),

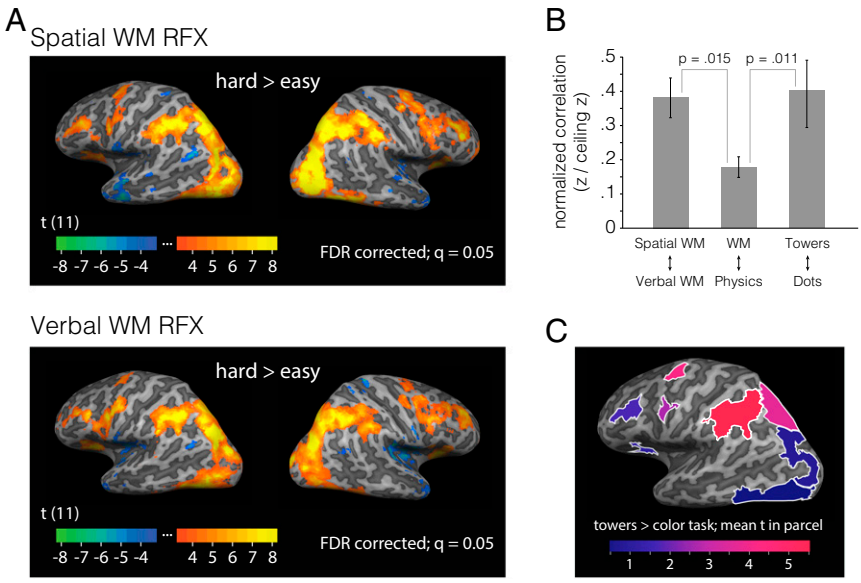

Fig. 5. Experiment 4 results: the relationship between physics-responsive brain regions and the MD network. (*A*) Group random effects (RFX) maps for the hard > easy contrast in the spatial working memory (spatial WM) and verbal working memory (verbal WM) tasks (task details are in Fig. S1). Voxels shown in the group RFX maps are significant after false discovery rate (FDR) correction at $q = 0.05$. (*B*) Correlation of the whole-brain pattern of blood-oxygen level-dependent response between pairs of tasks expressed as a proportion of the maximum possible correlation (*Materials and Methods*). Patterns of response for the spatial working memory and verbal working memory tasks were significantly more strongly correlated with each other than with the physics tasks. Likewise, patterns of response for the towers task and the dots task were significantly more strongly correlated with each other than with the working memory tasks. (*C*) Group parcels generated based on a hard > easy contrast in the spatial working memory and verbal working memory tasks shown for the left hemisphere. Parcels were generated by taking the intersection of the significant voxels for the two tasks within each subject using one-half of the data from each subject (run 2). The color of each parcel reflects the magnitude of the experiment 1 towers > color contrast within that parcel.

inherent interest (because the opposite contrast produced extensive activations in each case), or the spatial content of the physical tasks (which were matched in experiment 2). Neither can this pattern of response be explained by differential eye movements between conditions: within the candidate physical inference regions, the physics tasks produced stronger responses than a saccade task designed to elicit maximal eye movement-related responses (Figs. S2 and S3). Instead, this pattern of activation seems to reflect the process of physical scene understanding itself, which generalizes robustly across three tasks, each engaging different aspects of physical inference.

Although our data argue against the possibility that physical scene understanding is carried out by a purely domain-specific system, they also reject the possibility that physical inference is achieved by completely domain-general mechanisms (physical and nonphysical tasks were matched in difficulty, but we still found regions that were preferentially engaged by physical tasks). Instead, we find evidence for the third possible outcome proposed in the Introduction: brain regions exist that are preferentially engaged by physical inference over and above other similar and equally demanding scene understanding tasks (a physics engine in the brain), those regions are systematic across subjects, and they overlap with areas to which other functions have been previously attributed (namely motor action planning, tool use, and general problem-solving).

Does it make sense to talk about "the brain's physics engine" if these same regions are also engaged in other planning and problem-solving tasks? Consider an analogy with the GPUs that are now integrated into many computers. The highly parallel architecture of GPUs was originally motivated by the demands of graphics-intensive computing applications, but GPUs have since become indispensable for other applications, such as computer vision, deep neural network training, and indeed, real-time approximate physics simulation in computer games. An examination of the resource use of a computer would find that the GPU is active during all of these tasks and others that share similar computational demands. Thus, the same GPU hardware can serve as a physics engine, a graphics engine, a computer vision engine, and so forth—although it is not engaged by many other software applications, such as databases, word processors, or spreadsheets; it is not a completely general system, and it is not especially engaged in memory- or language-intensive processing. We propose that this analogy extends to the network of brain areas reported here, which are active for a set of tasks that shares similar computational demands and serves as a physics engine in the context of physically rich visual input or task demands.

What are those shared computational demands underlying physical inference, motor planning, and tool use that might lead to shared cortical systems? One possibility is that the ability to plan actions presumes a physical model of the world. Applying the correct force when grasping an object requires knowledge of the object's weight, its slipperiness, how much it will deform when grasped, etc. Clinical findings support this idea. Patients with acquired deficits in the ability to use familiar tools (apraxia) are sometimes also impaired in the ability to infer how a novel tool can be used based on its structure (essentially, an intuitive physics task) (47, 48). The loci of brain damage in such patients closely resemble the family of regions that we find to be engaged in physical inference. A second possible reason for the apparent overlap of physical inference activations with action planning/tool use regions is that we learn about the physical environment through interaction with it; learning about causality in infants is accelerated when they are enabled to engage in causal interventions at a younger age (49). Although it may be that physical inference necessarily involves some degree of covert action or motor imagery, we strove to minimize the motivation for motor imagery in experiments 2 and 3. Indeed, we saw the same set of action planning regions engaged when participants viewed highly simplified stimuli that had no 3D cues indicating how to interact with them (experiment 2) and when participants passively viewed movies and no action or judgment was required (experiment 3). Thus, a mental physics engine may be built into our brain's

action planning system (because action planning requires physical inference), but after we have this system in place, we may use it even when no action planning occurs.

Similar regions in premotor and parietal cortices have been shown to be engaged in spatial and temporal pattern prediction (50, 51) [in particular, in temporal order processing for abstract sequences (52)]. These abstract spatiotemporal predictions may be the building blocks on which the online mental simulation of physical interactions is built. As posited by Schubotz (50), the same motor circuits that calibrate our actions based on their predicted consequences (relying on feedback loops for online updating as an action unfolds) may be well-suited to performing online prediction of physical events, even in the absence of an action. The fact that physical inference may be rooted in an online updating mechanism agrees with the results of our experiments 2 and 3, where the action planning system was engaged simply by watching the physical events in a scene unfold. Importantly, however, our results argue against the idea that these regions are equally engaged in all types of spatiotemporal prediction—in experiment 2, we found that, even for difficulty-matched tasks that both involved spatial and temporal extrapolation of moving objects, the rules governing the motion (physical vs. social) significantly modulated the engagement of premotor and parietal cortices.

We have tested only a small subset of the full space of intuitive physical scene understanding. In daily life, we observe and predict not just the motions of rigid objects but also, the flow of fluids, the behavior of springs and pendulums, and the behavior of deformable objects, like ropes and cloth. Whether the observation and prediction of physical behaviors for these types of scenarios recruit the same brain regions that we found here remains to be seen, and the answer will speak to the scope and nature of the computations carried out in these regions. On one hand, all objects and materials are subject to the same physical laws and may be simulated by the same circuits that implement those laws. On the other hand, the actual behaviors of objects vary wildly depending on their particular properties—computationally effective procedures for simulating or judging collisions of rigid bodies do not apply at all for collisions of soft bodies. If the brain regions that we have uncovered for intuitive physical inference implement computations more like the simulation procedures of game physics engines [as hypothesized by Battaglia et al. (1)], rather than a set of universal physical laws, then we might expect to see activation patterns that vary with object or material type. Testing a broad range of intuitive physical inference tasks will be essential for understanding the function of each region and building models of how physical scene understanding is carried out in the brain.

This work opens up a broad landscape of new questions. Do the brain regions reported here explicitly code for physical properties, such as masses, forces, or materials? Do they code for events, such as collisions, falling, rolling, or sticking together? Do they interface with scene-selective cortical areas to establish a reference frame for physical predictions (53)? Some studies have found that premotor cortex encodes information about objects' weights (6, 7), but participants in these studies planned reaching movements to objects, where weight information is a necessary component of the motor plan to apply the correct force. It remains unknown whether the same regions represent mass and other physical properties in the absence of an intended action. Answering these questions may require examining how the multivariate pattern of response changes as a function of the physical content in a scene. Future work should also test the causal role of these regions in physical inference (for example, by intervening in the activity in these regions with transcranial magnetic stimulation). These studies and others can be expected to shed important new light on the fundamental everyday activity of physical scene understanding, its brain basis, and the computations that it entails.

Materials and Methods

The Massachusetts Institute of Technology (MIT) Institutional Review Boards approved all experimental protocols. All participants provided informed consent before their participation.

fMRI Data Collection and Preprocessing. fMRI data collection was conducted at the Athinoula A. Martinos Imaging Center at MIT on a Siemens 3T MAGNETOM Tim Trio Scanner (Siemens AG Healthcare) with a 32-channel head coil. A high-resolution T_1-weighted anatomical image (MPRAGE) was collected for each subject [repetition time (TR) = 2.53 s; echo time (TE) = 1.64, 3.5, 5.36, and 7.22 ms; flip angle-$\alpha = 7°$; field of view (FOV) = 256 mm; matrix = 256 × 256; slice thickness = 1 mm; 176 slices; acceleration factor = 3; 32 reference lines]. Functional data were collected using a T_2*-weighted echo planar imaging pulse sequence (TR = 2 s; TE = 30 ms; $\alpha = 90°$; FOV = 200 mm; matrix = 64 × 64; slice thickness = 3 mm; slice gap = 0.6 mm; 32 slices), which afforded whole-brain coverage.

Data preprocessing and general linear models were performed using a combination of BrainVoyager QX (Brain Innovation B.V.) and the FsFast tools in the FreeSurfer Software Suite (freesurfer.net). Surface visualizations were generated in BrainVoyager QX. All other analyses were conducted in Matlab R2012b (The MathWorks). Preprocessing consisted of 3D motion correction, slice scan time correction, high-pass filtering via a general linear model with a Fourier basis set (cutoff of two cycles per run, which also achieved linear trend removal), and spatial smoothing with a 4-mm FWHM Gaussian kernel. Before spatial smoothing, the functional runs were individually coregistered to the subject's T_1-weighted anatomical image. The data were transformed to a standardized coordinate system (Talairach space) to allow for group-level analyses. General linear models included 12 nuisance regressors based on the motion estimates generated from the 3D motion correction: *x*, *y*, and *z* translation; *x*, *y*, and *z* rotation; and the approximated first derivatives of each of these motion estimates.

Experiment 1. Thirteen subjects (ages 18–26 y old; six female) participated in the fMRI component of experiment 1. One participant was excluded (and the data were not analyzed), because the participant was too tired to complete the full experiment. All participants were right-handed and had normal or corrected to normal vision; 40 workers on AMT participated in the online portion of experiment 1. To participate in the task, AMT workers were required to have completed at least 1,000 previous human intelligence tasks (HITs) and have a 95% approval rating on previous HITs.

Stimuli and design. Stimuli for experiment 1 were based on those used by Battaglia et al. (1), and they were created in Blender 2.70 (Blender Foundation; https://www.blender.org). The stimuli were 6-s movies depicting stacks ("towers") of yellow, blue, and white blocks (Fig. 2*A*). The blocks were positioned so that the towers were unstable and would tumble if gravity was allowed to take effect. The facts that the towers would collapse and exactly how they would collapse were determined using rigid body physical simulation carried out by the Bullet Physics Engine in Blender. Each tower was positioned at the center of a circular floor; one-half of the floor was colored green, whereas the other one-half of the floor was colored red. Over the course of a 6-s movie, the camera viewpoint panned around the tower, completing one 360° pan, which allowed observers to see the tower from a range of vantage points. The tower itself was stationary within the scene. While viewing each movie, subjects were instructed to perform one of two tasks: (*i*) imagine how the blocks will fall and report whether more blocks will come to rest on the red or green side of the floor or (*ii*) report whether there are more blue or yellow blocks in the tower, ignoring any white blocks. Each tower contained a minimum of 17 blocks, not all of which were visible at a time, to discourage serial counting. Subjects were instructed to, instead, evaluate the proportion of the two colors in the tower as a whole. Each tower contained one or two more blocks of one color than the other, and some but not all towers contained one or two white blocks. The difficulty of the two tasks was matched based on data from AMT (*SI Materials and Methods*).

Each scanning run consisted of 23 18-s blocks: 10 blocks of the physical task, 10 blocks of the color task, and 3 rest blocks, which consisted of only a blank black screen. Each nonrest block began with a text cue, displayed for 1s, which read either "more blue or yellow?" (color task) or "where will it fall?" (physical task). These cues were chosen to be similar in length and avoid mention of color (which might prime attention to color) in the physical task cue. The text cue was followed by the presentation of a tower movie (6 s) and then, a black screen during a 2-s response period. This sequence was repeated twice within a block, with the same task being cued for both movie presentations within a block. Rest blocks occurred in blocks 1, 12, and 23, and the nonrest blocks were arranged in a pseudorandom palindromic order, so that the pairwise ordering between block types was balanced across

a run. A scanning run lasted for 414 s (207 volumes with a 2-s TR). Each subject participated in two runs collected in a random order.

Analysis. To identify group parcels and individual subject fROIs for the physics > color contrast, we used the group-constrained subject-specific ROI definition method (31, 32) on one-half of each subject's data (the second run that each subject completed). We thresholded individual subjects' significance maps at $P < 0.01$ uncorrected (an intentionally liberal threshold to facilitate identifying areas of group overlap) and identified voxels where five or more subjects showed significant activations. We then used a watershed algorithm to subdivide the group overlap map into discrete parcels. Individual subject fROIs were defined by intersecting the subject's thresholded map with each group-level parcel, and all subsequent analyses were performed in individual subject fROIs using the group parcels as an identification system to indicate corresponding ROIs across subjects.

Experiment 2. The same participants from the fMRI portion of experiment 1 participated in the fMRI portion of experiment 2. Additionally, 32 AMT workers participated in the online portion of experiment 2; to qualify to participate, workers were subject to the same criteria as in experiment 1.

Stimuli and design. The stimuli in experiment 2 were 10-s movies of red and blue dots moving within a square arena (Fig. 3*A*). The movement of the dots was dictated by either Newtonian mechanics for elastic collisions or social goals assigned to the dots [in the style of the classic animations by Heider and Simmel (33)]. *SI Materials and Methods* has details on the stimulus creation. Our goal was for subjects to generate expectations for how the dots would behave based on the physical or social behavior and then, predict how the dots would behave during a mental simulation period. For one-half of the movies in each set (balanced across the movies with barriers or no barriers), we generated a version of the video in which one of the dots became invisible during the last 2 s of the 10-s movie and became visible again on the final frame of the movie. The invisible dot still interacted either socially or physically with the other dot and the arena. These movies were labeled "correct" movies, where the final position of the hidden dot, when it reappeared on the last frame, was consistent with where it would have traveled given its behavior before becoming invisible. We then generated a second set of movies, labeled "incorrect" movies, where we displaced the final position of the hidden dot when it reappeared, so that the final position was inconsistent with the behavior of the dot before its disappearance. A participant's task when viewing each movie was to observe the dot behavior, imagine how the hidden dot was behaving during the 2 s when it was invisible, and then, report whether the final position of the dot when it reappeared was correct or incorrect. We matched the difficulty of the judgments for the social and physical movies based on data collected from AMT (*SI Materials and Methods*).

Each scanning run consisted of 19 26-s blocks: 8 blocks containing physical movies, 8 blocks containing social movies, and 3 rest blocks, which consisted of only a blank black screen. Each nonrest block contained two movies of the same type (physical or social). A block began with the presentation of a 10-s movie, and when the movie was finished, the final frame stayed onscreen for another 1.5 s (during which subjects decided whether the final dot position was correct and made a response). A blank screen was displayed for 1.5 s before the onset of the next movie. This sequence was repeated twice within a block, amounting to a 26-s block in total. Rest blocks occurred in blocks 1, 10, and 19, and the nonrest blocks were arranged in a palindromic order as in experiment 1. A scanning run lasted for 494 s (247 volumes with a 2-s TR). Each subject participated in two runs, collected in a random order.

Analysis. Data were preprocessed and aligned to each subject's high-resolution anatomical image as in experiment 1. We analyzed the data within the candidate physics-responsive regions obtained with the group-constrained subject-specific method in experiment 1. Because the ROIs were defined independently on the data from experiment 2, we used both of a subject's runs to generate the percentage signal change (PSC) plots shown in Fig. 3 *B* and *C*. All other details of the analysis are identical to those in experiment 1.

Experiment 3. Sixty-five subjects participated in the fMRI component of experiment 3. Data from a subset of these participants were previously published in the work by Julian et al. (31). Additionally, 32 AMT workers participated in the online portion of experiment 3; to qualify to participate, workers were subject to the same criteria as in experiments 1 and 2.

Stimuli and design. Details of the stimuli and design of experiment 3 are published in the work by Julian et al. (31) In brief, participants were presented with 3-s movie clips of faces, bodies, scenes, objects, and scrambled objects. Face and body movies showed children playing in front of a black background, scene movies showed various environments (mostly pastoral scenes) passing by as the camera moved through them, and object movies showed toys, such as balls or vehicles, in motion (e.g., spinning or rolling). Scrambled objects movies were created by dividing each object movie into a 15×15 grid and randomly rearranging the locations of the cells in the grid. There were 60 movies in each category (300 movies total), and the movies presented from each category during a run were randomly sampled without replacement from the full set of available movies. Movie clips were organized into blocks that contained six clips from the same category (18-s blocks), and each scanning run contained two blocks from each category. A run also included three 18-s rest blocks (blocks 1, 7, and 13), during which a full-screen color was displayed, and the color changed every 3 s. Subjects were simply instructed to watch the stimuli that appeared onscreen. Each subject completed four 234-s runs of the experiment. Acquisition parameters for fMRI data collection differed slightly from those used in experiments 1 and 2 but still afforded full-brain coverage (TR = 2,000 ms; TE = 30 ms; FOV = 192×192; matrix = 64×64; 32 slices; $3 \times 3 \times 3.6$-mm voxel resolution). We obtained ratings of the perceived physical content in the videos from each category from workers on AMT (*SI Materials and Methods*).

Analysis. We identified group parcels and individual subject fROIs for the objects > scrambled objects contrast using the same approach as in experiment 1, using one-half of each subject's data (runs 2 and 4). We used the independent, left-out data from each subject (runs 1 and 3) to generate PSC plots for five stimulus categories (Fig. 4*C*).

ACKNOWLEDGMENTS. We thank Ray Gonzalez for assistance with web programming and data collection, Dr. Marina Bedny for comments on an earlier draft of the manuscript, and Dr. Daniel Dilks for helpful discussions. This work was supported by Eunice Kennedy Shriver National Institute of Child Health and Human Development Award F32-HD075427 (to J.F.), National Eye Institute Grant EY13455 (to N.K.), and National Science Foundation Science and Technology Center for Brains, Minds, and Machines Grant CCF-1231216.

1. Battaglia PW, Hamrick JB, Tenenbaum JB (2013) Simulation as an engine of physical scene understanding. *Proc Natl Acad Sci USA* 110(45):18327–18332.
2. Cant JS, Arnott SR, Goodale MA (2009) fMR-adaptation reveals separate processing regions for the perception of form and texture in the human ventral stream. *Exp Brain Res* 192(3):391–405.
3. Newman SD, Klatzky RL, Lederman SJ, Just MA (2005) Imagining material versus geometric properties of objects: An fMRI study. *Brain Res Cogn Brain Res* 23(2-3): 235–246.
4. Goda N, Tachibana A, Okazawa G, Komatsu H (2014) Representation of the material properties of objects in the visual cortex of nonhuman primates. *J Neurosci* 34(7): 2660–2673.
5. Gallivan JP, Cant JS, Goodale MA, Flanagan JR (2014) Representation of object weight in human ventral visual cortex. *Curr Biol* 24(16):1866–1873.
6. Loh MN, Kirsch L, Rothwell JC, Lemon RN, Davare M (2010) Information about the weight of grasped objects from vision and internal models interacts within the primary motor cortex. *J Neurosci* 30(20):6984–6990.
7. van Nuenen BF, Kuhtz-Buschbeck J, Schulz C, Bloem BR, Siebner HR (2012) Weight-specific anticipatory coding of grip force in human dorsal premotor cortex. *J Neurosci* 32(15):5272–5283.
8. Mason RA, Just MA (2016) Neural representations of physics concepts. *Psychol Sci* 27(6):904–913.
9. Jack AI, et al. (2013) fMRI reveals reciprocal inhibition between social and physical cognitive domains. *Neuroimage* 66:385–401.
10. Kestenbaum R, Termine N, Spelke ES (1987) Perception of objects and object boundaries by 3-month-old infants. *Br J Dev Psychol* 5(4):367–383.
11. Spelke ES, Kestenbaum R, Simons DJ, Wein D (1995) Spatiotemporal continuity, smoothness of motion and object identity in infancy. *Br J Dev Psychol* 13(2):113–142.
12. Hespos SJ, Ferry AL, Rips LJ (2009) Five-month-old infants have different expectations for solids and liquids. *Psychol Sci* 20(5):603–611.
13. Hespos SJ, Ferry AL, Anderson EM, Hollenbeck EN, Rips LJ (2016) Five-month-old infants have general knowledge of how nonsolid substances behave and interact. *Psychol Sci* 27(2):244–256.
14. Leslie AM, Keeble S (1987) Do six-month-old infants perceive causality? *Cognition* 25(3):265–288.
15. Oakes LM (1994) Development of infants' use of continuity cues in their perception of causality. *Dev Psychol* 30(6):869–879.
16. Needham A, Baillargeon R (1997) Object segregation in 8-month-old infants. *Cognition* 62(2):121–149.
17. Baillargeon R (1998) Infants' understanding of the physical world. *Advances in Psychological Science, Vol. 2: Biological and Cognitive Aspects*, eds Sabourin M, Craik F, Robert M (Psychology Press/Erlbaum UK Taylor & Francis, Hove, England), pp 503–529.
18. Leslie AM (1995) A theory of agency. *Causal Cogn Multidiscip Debate*, eds Premack AJ, Premack D, Sperber D (Clarendon Press, Oxford), pp 131–149.
19. Spelke ES, Breinlinger K, Macomber J, Jacobson K (1992) Origins of knowledge. *Psychol Rev* 99(4):605–632.

20. McCloskey M, Caramazza A, Green B (1980) Curvilinear motion in the absence of external forces: Naive beliefs about the motion of objects. *Science* 210(4474): 1139–1141.
21. Caramazza A, McCloskey M, Green B (1981) Naive beliefs in "sophisticated' subjects: Misconceptions about trajectories of objects. *Cognition* 9(2):117–123.
22. McCloskey M, Washburn A, Felch L (1983) Intuitive physics: The straight-down belief and its origin. *J Exp Psychol Learn Mem Cogn* 9(4):636–649.
23. Smith K, Battaglia P, Vul E (2013) Consistent physics underlying ballistic motion prediction. *Proceedings of the 35th Conference of the Cognitive Science Society, Berlin, Germany, July 31-August 3, 2013*, eds Knauff M, Pauen M, Sebanz N, Wachsmuth I (Cognitive Science Society, Austin, TX), pp 3426–3431.
24. Smith KA, Vul E (2013) Sources of uncertainty in intuitive physics. *Top Cogn Sci* 5(1): 185–199.
25. Sanborn AN, Mansinghka VK, Griffiths TL (2013) Reconciling intuitive physics and Newtonian mechanics for colliding objects. *Psychol Rev* 120(2):411–437.
26. Duncan J, Owen AM (2000) Common regions of the human frontal lobe recruited by diverse cognitive demands. *Trends Neurosci* 23(10):475–483.
27. Fedorenko E, Duncan J, Kanwisher N (2013) Broad domain generality in focal regions of frontal and parietal cortex. *Proc Natl Acad Sci USA* 110(41):16616–16621.
28. Kanwisher N, McDermott J, Chun MM (1997) The fusiform face area: A module in human extrastriate cortex specialized for face perception. *J Neurosci* 17(11): 4302–4311.
29. Overath T, McDermott JH, Zarate JM, Poeppel D (2015) The cortical analysis of speech-specific temporal structure revealed by responses to sound quilts. *Nat Neurosci* 18(6): 903–911.
30. Saxe R, Kanwisher N (2003) People thinking about thinking people. The role of the temporo-parietal junction in "theory of mind." *Neuroimage* 19(4):1835–1842.
31. Julian JB, Fedorenko E, Webster J, Kanwisher N (2012) An algorithmic method for functionally defining regions of interest in the ventral visual pathway. *Neuroimage* 60(4):2357–2364.
32. 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.
33. Heider F, Simmel M (1944) An experimental study of apparent behavior. *Am J Psychol* 57(2):243–259.
34. Decety J, Grèzes J (1999) Neural mechanisms subserving the perception of human actions. *Trends Cogn Sci* 3(5):172–178.
35. Wende KC, et al. (2013) Differences and commonalities in the judgment of causality in physical and social contexts: An fMRI study. *Neuropsychologia* 51(13):2572–2580.
36. Mordatch I, Todorov E, Popović Z (2012) Discovery of complex behaviors through contact-invariant optimization. *ACM Trans Graph* 31(4):43.
37. Graziano MS, Botvinick MM (2002) How the brain represents the body: Insights from neurophysiology and psychology. *Common mechanisms in perception and action: Attention and performance XIX*, eds Prinz W, Hommel B (University Press, Oxford), pp 136–157.
38. Duncan J (2010) The multiple-demand (MD) system of the primate brain: Mental programs for intelligent behaviour. *Trends Cogn Sci* 14(4):172–179.
39. Connolly JD, Andersen RA, Goodale MA (2003) FMRI evidence for a 'parietal reach region' in the human brain. *Exp Brain Res* 153(2):140–145.
40. Gallivan JP, Culham JC (2015) Neural coding within human brain areas involved in actions. *Curr Opin Neurobiol* 33:141–149.
41. Gazzola V, Keysers C (2009) The observation and execution of actions share motor and somatosensory voxels in all tested subjects: Single-subject analyses of unsmoothed fMRI data. *Cereb Cortex* 19(6):1239–1255.
42. Simon SR, et al. (2002) Spatial attention and memory versus motor preparation: Premotor cortex involvement as revealed by fMRI. *J Neurophysiol* 88(4):2047–2057.
43. Tanji J, Shima K (1994) Role for supplementary motor area cells in planning several movements ahead. *Nature* 371(6496):413–416.
44. Gallivan JP, McLean DA, Valyear KF, Culham JC (2013) Decoding the neural mechanisms of human tool use. *eLife* 2:e00425.
45. Brandi M-L, Wohlschläger A, Sorg C, Hermsdörfer J (2014) The neural correlates of planning and executing actual tool use. *J Neurosci* 34(39):13183–13194.
46. Valyear KF, Gallivan JP, McLean DA, Culham JC (2012) fMRI repetition suppression for familiar but not arbitrary actions with tools. *J Neurosci* 32(12):4247–4259.
47. Goldenberg G, Hagmann S (1998) Tool use and mechanical problem solving in apraxia. *Neuropsychologia* 36(7):581–589.
48. Goldenberg G, Spatt J (2009) The neural basis of tool use. *Brain* 132(Pt 6):1645–1655.
49. Rakison DH, Krogh L (2012) Does causal action facilitate causal perception in infants younger than 6 months of age? *Dev Sci* 15(1):43–53.
50. Schubotz RI (2007) Prediction of external events with our motor system: Towards a new framework. *Trends Cogn Sci* 11(5):211–218.
51. Schubotz RI (2004) *Human Premotor Cortex: Beyond Motor Performance. Dissertation (Max Planck Institute for Human Cognitive and Brain Sciences Leipzig)*. Available at pubman.mpdl.mpg.de/pubman/faces/viewItemOverviewPage.jsp?itemId=escidoc:723119:3. Accessed October 17, 2015.
52. Schubotz RI, von Cramon DY (2004) Sequences of abstract nonbiological stimuli share ventral premotor cortex with action observation and imagery. *J Neurosci* 24(24): 5467–5474.
53. Vaziri S, Connor CE (2016) Representation of gravity-aligned scene structure in ventral pathway visual cortex. *Curr Biol* 26(6):766–774.
54. Amiez C, Kostopoulos P, Champod A-S, Petrides M (2006) Local morphology predicts functional organization of the dorsal premotor region in the human brain. *J Neurosci* 26(10):2724–2731.

Supporting Information

Fischer et al. 10.1073/pnas.1610344113

SI Materials and Methods

Data Collection from Amazon Mechanical Turk.

Experiment 1. We began with an initial set of 150 movies of block towers and had 40 workers on AMT make the physical and color judgments described in *Materials and Methods*. Each worker made judgments on the complete set of movies. The data from three workers were excluded from additional analysis after it was determined that they ignored the task instructions (two workers provided the same response for all movies, and one provided random responses). From the full set of movies, we selected a subset of 40 that had the lowest within-movie difference in percentage correct between the physical and color judgments subject to counterbalancing constraint that the selected set contained equal numbers of movies, in which the correct responses for the physical and color judgments were red, blue; red, yellow; green, blue; and green; yellow. This set of 40 movies was divided into sets for two functional runs, each adhering to the same counterbalancing constraint. The accuracies within the movies selected for each run were set 1: physical task, 83.8 ± 12.1 SD percentage correct; color task, 83.8 ± 13.8 SD percentage correct; $t(19) = 0$ for a difference between tasks; $P = 1$; paired t test and set 2: physical task, 87.0 ± 12.6 SD percentage correct; color task, 86.4 ± 12.2 SD percentage correct; $t(19) = 0.26$ for a difference between tasks; $P = 0.79$; paired t test. Each individual movie appeared twice within a functional run: once with instructions to perform the physical task and once with instructions to perform the color task. In this way, we equated the difficulty of the physical simulation and color judgment tasks within each functional run to minimize the possibility that any difference in mental effort between the conditions could give rise to differences in the fMRI BOLD response.

Experiment 2. To match the difficulty of the judgments for the social and physical movies, we collected data from 32 workers on AMT. Each worker performed the correct/incorrect position judgment for all 80 movies and also, provided an "animacy" rating for each movie. The instruction for the animacy rating was: "For each movie, we will ask you how animated the circles looked—in other words, to what extent did the circles look like they were moving on their own as a person would vs. moving as an inanimate billiard ball would?" AMT workers rated the animacy of each movie on a scale from one (not alive) to five (strongly appeared to be alive). The data from two workers were excluded from additional analysis because of quality issues (one provided the same response to nearly every movie and performed at chance; the other consistently selected a response before seeing the final ball position). The ratings confirmed a significant difference in the perceived animacy of the dots in the physical vs. social movies [mean rating of 1.62 ± 0.25 SD for physical movies; mean rating of 3.85 ± 0.45 SD for social movies; $t(78) = 27.3$; $P = 10^{-42}$; two-sample t test]. From the original set of 80 movies, we retained 64 movies, which allowed for a good match in percentage correct between the physical and social judgments. These movies were divided into two runs of 32 movies each (16 physical and 16 social movies), with one-half of the movies in each run containing barriers and the other one-half containing no barriers. We also ensured that the dot that disappeared in each movie was red or blue with equal frequency to discourage preferential attention to one color. The accuracies within the movies selected for each run were set 1: physical movies, 75.0 ± 18.4 SD percentage correct; social movies, 72.7 ± 16.7 SD percentage correct; $t(30) = 0.38$ for a difference between tasks; $P = 0.71$; two-sample t test and set 2: physical movies, 78.4 ± 18.1 SD percentage correct; social movies, 78.1 ± 12.9 SD percentage correct; $t(30) = 0.047$ for a difference between tasks; $P = 0.96$; two-sample t test.

Experiment 3. To test for differences in the perceived physical content among five categories, we collected ratings of the movies from 32 workers on AMT. Workers viewed all 300 movies in a random order and responded to the following question: "When viewing the video, to what extent do you notice and focus on the physical properties of the objects or people in the clip (for example, their weight, texture, or hardness), and to what extent do you notice and focus on the physical interactions between the objects or people (for example, contact, collision, rolling, or tumbling)?" Workers provided a response on a scale from one (least physical content) to five (most physical content) for each movie. In addition, workers performed a one-back task included to monitor their attentiveness—they reported whether each movie was identical to the movie that they had seen on the previous trial. Of 32 workers who originally participated, one worker's data were excluded from additional analysis because of chance performance on the one-back task, and a second worker's data were excluded for providing the same response on nearly every video. The physical content ratings are reported in Fig. 4*A*.

Experiment 2 Stimulus Creation. Each dot in the physical and social movies had a radius of 25 pixels, and the full arena was 200 × 200 pixels. Some arenas had gray barriers (rectangles) of various sizes that the dots could not pass through. To generate movies with physical behavior, we assigned starting positions and velocities to the dots and then, computed their updated positions in fine temporal increments. The updated 2D position P of a dot at time t was given by

$$P_{t+1} = P_t + V_t \times dt,$$

where V was a dot's 2D velocity vector, and dt was the temporal sampling increment of the simulation ($10^{-}5$). The change in V caused by rolling friction was computed as

$$V_{t+1} = V_t \times \frac{1 - \mu_0 \times dt}{\sqrt{V_{t,x}^2 + V_{t,y}^2}},$$

where μ_0 was a constant ($\mu_0 = 500$) determining the loss of energy because of friction with the arena floor. A collision of a dot with a wall or another dot similarly incurred an energy loss given by

$$V_t^{'} = V_t \times \frac{1 - \mu_1}{\sqrt{V_{t,x}^2 + V_{t,y}^2}},$$

where $\mu_1 = 1.5$. A wall collision resulted in an inversion of the appropriate x or y component of the velocity. When the two dots collided with each other, they exchanged their velocities along the normal line connecting their centers (implying equal masses for the two dots). One frame for every 800 time steps was saved to an image, and the resulting images were assembled into a 10-s movie. The resulting movies resembled billiard balls or hockey pucks bouncing within the arena.

To generate the movies with social behavior, we animated the dot motion by hand. For each movie, we began with the same barriers and starting positions as one of the physical movies and

had in mind a simple narrative for the interaction of the dots (e.g., blue is chasing red, and red wants to hide in the corner). Using custom Matlab scripts, we controlled the dots using a computer trackpad and recorded their positions on a frame by frame basis. The dots were not allowed to pass through walls, barriers, or each other, and they were animated to generally avoid collisions so as to reduce the physical content in the movies. Using this approach, we generated 40 physical movies and 40 social movies that were paired in terms of the arena arrangement and dot starting positions. One-half of the movies in each set had barriers positioned within the arena, and one-half did not.

Bootstrapping and Statistics. Error bars for all plots were generated using a bootstrapping approach. The effect of interest was first computed within each subject. Then, on each of 10,000 bootstrap iterations, the subjects in the group were sampled with replacement, and the group mean was computed on each iteration. Error bars displayed are the SDs of the upper and lower halves of the bootstrapped distribution (computed separately), which correspond to the SEM.

PSC plots were generated by averaging the signal time course across voxels in an ROI and then, extracting the time points for the rest conditions. We used the mean signal from the last four time points within the rest blocks as a baseline for computing PSC (excluding earlier time points reduced the amount of signal carryover from previous blocks that were included in the baseline). We then computed PSC for the remaining time points by dividing their intensity values by the baseline, multiplying by 100, and subtracting 100. To plot the PSC for a condition, we collected all of the blocks for that condition, averaged the PSC across blocks within a subject, and computed error bars across subjects using a bootstrapping method as described above. For significance testing and displaying bar plots, we removed the first three time points of the PSC time course (to accommodate the rise time of the BOLD response after the onset of a block) and averaged the PSC across the rest of the block. Error bars for bar plots were computed across subjects with the same bootstrapping approach, and significance testing for differences between bars was conducted with a permutation analyses as described above.

MD Network Localizer. We identified the MD network (26, 38) in the same subjects who participated in experiments 1 and 2 using two working memory tasks adapted from the work by Fedorenko et al. (27). Details of the tasks are provided in Fig. S1. First-level analyses of the data were conducted in the same way as in experiments 1 and 2. To compute the correlation between the patterns of response for the physics tasks and the working memory tasks, we split each subject's data from each of the tasks in half, generating BOLD response maps for each individual run in each subject. The spatial and verbal working memory maps (spatial_WM and verbal_WM, respectively) reflected a hard > easy contrast, and the physics maps reflected a physics > color contrast (experiment 1) and a physical > social contrast (experiment 2). Because the negative responses in the physics contrasts reflected different types of information by design and would be expected to appear in different brain areas (e.g., color-responsive regions in experiment 1 and social processing regions in experiment 2), we removed the negative responses from all maps and performed correlations only on the positive responses. We estimated the split-half reliability for the maps from each task by taking the two maps for a given task within a subject (originating from two runs) and computing the correlation between them. This within-task correlation provided an estimate of the largest possible correlation that we could expect when correlating one of the maps in the pair with a map coming from a different task based on the amount of noise in the individual maps. We then computed all of the pairwise correlations among the individual run maps for spatial working memory, verbal working memory, towers, and physical dots within each subject. We normalized each of these pairwise correlations by the maximum possible correlation for that pair given by the within-task correlations. To combine the two within-task correlations for a pair into a single estimate of the maximum possible correlation for the pair of tasks, we used a Monte Carlo simulation. On each of 10,000 iterations, a random signal of the same length as the fMRI data was generated to use as a source signal. Multiple copies of the source signal were then corrupted with varying amounts of noise, and we selected the appropriate noise levels to produce the within-task correlations found for the two tasks in the pair. We then computed the across-task correlation, which gave the ceiling correlation used to normalize the empirical correlation between the pair of tasks. Because the same underlying source signal was used to generate the noise-corrupted signals corresponding to the two tasks, the correlation between tasks reflected the maximum possible correlation (i.e., the correlation if the two tasks actually originated from the same underlying signal). Correlations shown in Fig. 5*B* are the normalized correlations averaged across runs (e.g., the spatial_WM↔verbal_WM correlation is the average of $\text{spatial\_WM}_{\text{run\_1}}$↔$\text{verbal\_WM}_{\text{run\_1}}$, $\text{spatial\_WM}_{\text{run\_1}}$↔$\text{verbal\_WM}_{\text{run\_2}}$, $\text{spatial\_WM}_{\text{run\_2}}$↔$\text{verbal\_WM}_{\text{run\_1}}$, and $\text{spatial\_WM}_{\text{run\_2}}$↔$\text{verbal\_WM}_{\text{run\_2}}$).

To generate MD network parcels, we used the same process as in experiment 1, except that we used the intersection of the significant voxels from the spatial working memory hard > easy contrast and the verbal working memory hard > easy contrast to create each subject's significance map, which then contributed to the group overlap map.

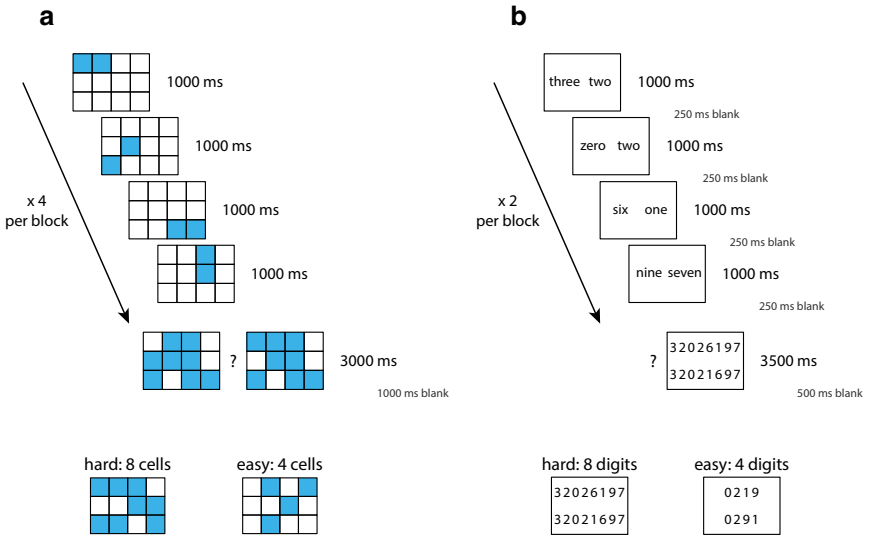

Fig. S1. Spatial and verbal working memory tasks. We localized the MD network (26, 38) using two working memory tasks adapted from the work by Fedorenko et al. (27). (*A*) Spatial working memory task. On each trial, participants saw a series of four 3 × 4 grids presented for 1 s each. In each grid, either one or two of the cells were filled in blue (one cell per grid for the "easy" condition and two cells per grid for the "hard" condition). Participants had to remember the locations of the filled cells. During a subsequent 3-s response period, two grids were displayed side by side; one contained the correct summation of the filled locations from the previous four grids, and one contained an incorrect arrangement generated by shifting one of the filled cells to an incorrect location. Participants responded, indicating which of the grids contained the correct arrangement. A block comprised four trials of the same condition (32-s blocks), and a run contained 12 blocks plus four 16-s rest periods (448 s per run). Each subject completed two runs of the task. (*B*) Verbal working memory task. Participants saw four frames displayed for 1 s each, which contained written digits (one digit per frame for the easy condition and two digits per frame for the hard condition). Participants had to remember the sequence of digits presented. On a final frame displayed for 3.5 s, two digit strings were presented: one containing the correct ordering of the digits from the previous four frames and one containing an incorrect ordering generated by swapping two of the digits in the sequence. Participants selected the string with the correct ordering. A block contained two trials of the same condition (18 s), and a run contained 20 blocks plus four 18-s rest periods (432 s total for a run). Each subject participated in two runs.

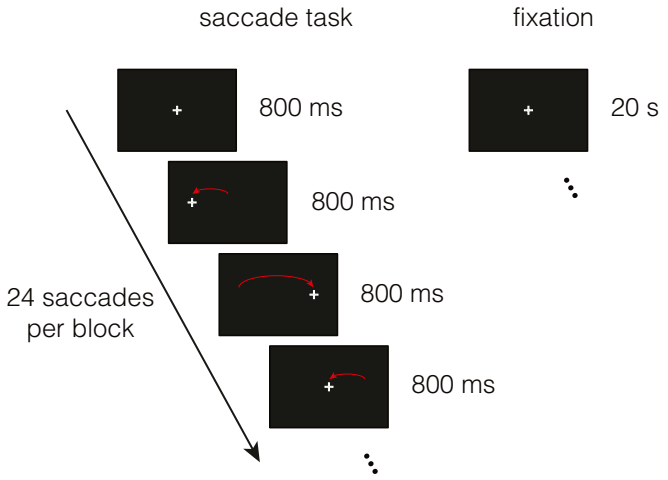

Fig. S2. Eye movement task. To identify responses related to eye movements, we used a saccade task adapted from the work by Amiez et al. (54). Participants were cued to perform the saccade task with a 2-s text cue ("follow the fixation cross") and then, saw a 1-s blank screen. Subjects then saw a series of 800-ms frames, in which the fixation cross appeared in one of three locations: the center of the screen or 8.4° to the left or right of center. The position of the fixation cross was chosen randomly on each frame subject to the constraint that it appeared eight times in each of the left and right sides of the screen and nine times in the center over the course of 25 frames. Subjects had to make a saccade to the new location of the fixation cross as quickly as possible each time that its location changed. A block ended with a 3-s period when a static fixation cross was displayed in the center of the screen, making 26-s blocks. In separate rest blocks, also 26 s, the text cue read "fixate the cross," and the fixation cross remained in the center of the screen for the duration of the block. Four blocks each of the saccade condition and the rest condition were presented within a run. Runs also contained blocks of another motor localizer task not analyzed for this study. A full run lasted 416 s. Five subjects who also participated in experiments 1 and 2 participated in the eye movement task, and each subject completed two runs.

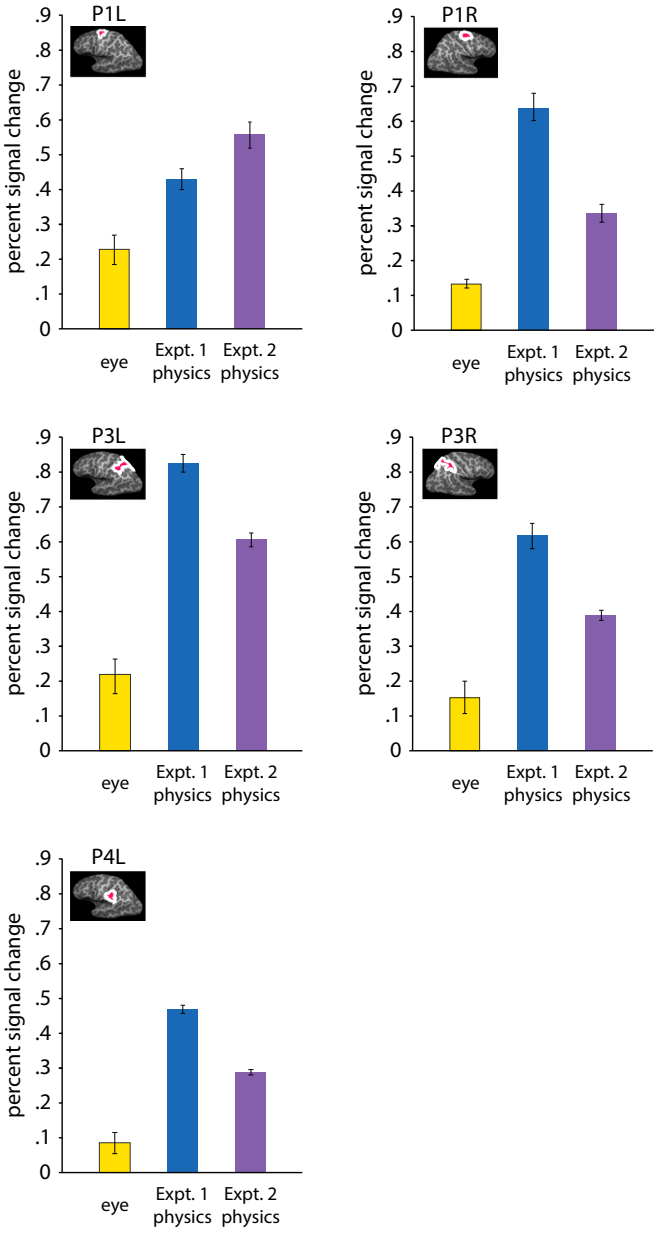

Fig. S3. Eye movements cannot account for responses within the candidate physics regions. Within five candidate physics regions identified in experiments 1 and 2, we compared the responses to the physics tasks with those from the eye movement task described in Fig. S2. PSC for the three tasks is shown here, and it was computed based on five subjects who completed the eye movement task. Although eye movement responses were significant in these areas ($t_4 = 6.67$; $P = 0.0026$; one-sample t test on eye movement PSCs averaged across areas), the response to the physical inference conditions from experiments 1 and 2 was significantly larger than that to saccades ($t_4 = 7.65$; $P = 0.0016$ for experiment 1 physics vs. eye movements; $t_4 = 10.61$; $P = 0.00045$ for experiment 2 physics vs. eye movements; paired t tests on PSC values averaged across five areas). These data rule out the possibility that any differential eye movements between conditions in the physics tasks could have driven the physics-related responses: the physics stimuli produced larger responses than even a saccade-based localizer designed specifically to maximize the BOLD response related to eye movements.