NEUROSCIENCE

Decoding predicted future states from the brain's "physics engine"

R. T. Pramod[1,2]*, Elizabeth Mieczkowski[1,2,3], Cyn X. Fang[1,2], Joshua B. Tenenbaum[1,2], Nancy Kanwisher[1,2]

Successful engagement with the physical world requires the ability to predict future events and plan interventions to alter that future. Growing evidence implicates a set of regions in the human parietal and frontal lobes [also known as the "physics network" (PN)] in such intuitive physical inferences. However, the central tenet of this hypothesis, that PN runs forward simulations to predict future states, remains untested. In this preregistered study, we first show that PN abstractly represents whether two objects are in contact with each other, a physical scene property critical for prediction (because objects' fates are intertwined when they are in contact). We then show that PN (but not other visual areas) carries abstract information about predicted future contact events (i.e., collisions). These findings support the hypothesis that PN contains a generative model of the physical world that conducts forward simulations, serving as the brain's "physics engine."

INTRODUCTION

To plan even the most mundane action, we must predict future states of the world. Catching a ball requires predicting its trajectory, placing an object stably on a surface requires predicting whether it will fall, and changing lanes in traffic requires predicting where cars in the adjacent lane will be in a few seconds. How do we make these predictions? All prediction requires prior knowledge about how the world works, enabling us to generate likely future states from the present estimated state. Here, we focus on the case of predicting physical events. Specifically, we test the hypothesis that a set of brain regions previously implicated in intuitive physical reasoning spontaneously generates predictions of future states when we simply view a short video of objects in motion, even without any explicit prediction or planning task.

The hypothesized "physics network" (PN) includes a set of bilateral parietal and frontal regions that were first identified in functional magnetic resonance imaging (fMRI) studies as responding more strongly when people perform simple physical reasoning tasks (which way will the tower fall?) than perceptual judgments on the same stimuli (does the tower contain more yellow or blue blocks?) (*1*). PN was subsequently shown to carry information about object mass (*2*) and physical stability (*3*) that generalized across scenarios, and to show an increased response when physical, but not social, expectations are violated (*4*). These results provide initial evidence for the physics engine hypothesis—that PN contains a generative model of the physical world capable of running forward simulations. However, prior work has not provided explicit evidence that predicted future states are represented in PN before they occur. In the current study, we use the case of object contact relationships to directly probe for future prediction in PN.

Object contact relationships such as containment, support, and attachment are critical for dynamic physical scene understanding and for predicting what will happen next. When two objects are in contact, their fate is intertwined: If a container moves, then so does its containee, but the same is not true of an object that is merely occluded by the container without contacting it (*5*). Befitting their fundamental importance for understanding the physical world, object contact relations are privileged in language (e.g., "in" versus "on" are included in the closed class of spatial prepositions), adult perception [where contact relations are extracted quickly and arguably automatically (*6*)], and development. Infants as young as 3 months expect objects to move with other objects that contain them but not with objects that merely occlude them without contact (*7*). Infants are also sensitive to support relationships: Around 6.5 months, they understand that more than half of the base of the supported object must be in contact with the supporting object, and later at around 12.5 months, they begin to take into account the supported object's shape or proportional distribution to infer stable support (*8*). Last, infants (6 and 7 months) also look longer if an object moves without having been contacted by another moving object (*9*), again highlighting the fundamental role of object contact in intuitive physical reasoning. Given their importance for physical prediction, we hypothesized that PN would encode contact relationships between objects. We test this hypothesis in experiment 1.

We then use that finding to test a central tenet of the "physics engine" hypothesis, that we run forward simulations to predict what will happen next (*10*). If PN is the physical instantiation of a mental physics engine, then it should contain information about predicted future states before they occur. To test this hypothesis, in experiment 2, we scan participants while they view contact events (collisions) and no-contact events (noncollisions), and we ask whether the neural distinction between perceived collision/noncollision is also found for predicted collision/noncollision (i.e., when participants view videos in which impending collisions can be predicted but are not observed). Cross-decoding of collision/noncollision from the perceived to the predicted case would support the idea that PN represents predicted future states.

A third question addressed in this work concerns the level of abstraction in the representation of contact percepts and predictions. A physics engine might be expected to represent abstract object relations such as "contact" independently from the representation of the objects involved, perhaps even generalizing to some extent across contact types (e.g., support, containment, and attachment).

[1]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02138, USA. [2]Center for Brains, Minds and Machines, Cambridge, MA 02138, USA. [3]Department of Computer Science, Princeton University, Princeton, NJ 08544, USA.
*Corresponding author. Email: pramodrt@mit.edu

Although such abstractions are not sufficient for precise simulations of what will happen next, they are still useful for simplifying the computations required for forward simulation. If a potentially movable object is not currently moving and not in contact with other moving objects, then the physics engine need not update the state of that object [relegating it to a "sleep" state (*11*)] until the object makes a new contact relationship. If a movable object is supported by, contained within, or attached to another moving object, then that contact qualitatively constrains its motion (at least locally) and reduces the number of degrees of freedom that must be tracked and updated in the dynamic scene. Thus, abstractions can be computationally efficient and can also serve as bridges for relating prior experiences to the current scenario. We test the abstractness of the brain's representations of current and predicted contact by testing whether we can cross-decode from PN the current or future presence versus absence of object contact, across object identities, shapes, configurations, and motion trajectories. Of course, more fine-grained representations of specific object shapes, configurations, and trajectories will also be required for the physical simulations that have been posited to underlie many aspects of prediction and planning (*10*, *12*). As in hierarchical approaches to task and motion planning in robotics (*13*), the brain could use a hierarchy of models for planning, where more abstract simulations of dynamics support high-level goal-based decomposition of a task into subgoals and subtasks—that is, an abstract plan to achieve a goal—while more fine-grained simulations are used to predict and generate the precise motions and motor action sequences needed to implement this plan in a specific task setting (see the Discussion section for more).

RESULTS

Experiment 1: Scenario-invariant decoding of object contact relationship in the hypothesized PN

In experiment 1, we used both naturalistic and artificially rendered videos (Fig. 1; see the Materials and Methods for stimulus design) depicting various object relationships to test whether the PN represents object contact relationships. We hypothesized that any brain region involved in physical reasoning should represent the presence of object-object contact because contact constrains how objects move and thus is critical for prediction. We first identified functional regions of interest (fROIs) in each participant individually, including the PN in the frontoparietal cortices, the lateral occipital complex (LOC), and the ventral temporal cortex (VTC) (see Materials and Methods). We then collected fMRI responses for each voxel in these fROIs to each of the contact and noncontact relationships across three different scenarios (natural-create, natural-consequence, and rendered; see Fig. 1). These responses were used to conduct multivoxel pattern analyses (MVPAs) to test for the presence of object contact information both in PN and in other cortical regions.

Contact versus noncontact decoding

We used correlation-based MVPA to test whether contact relationships can be distinguished from noncontact relationships, invariant to the underlying scenario, within each fROI (Fig. 2A; see the Materials and Methods). We quantified the presence of contact information within an fROI using a decoding index: the correlation of the pattern of response across voxels within contact relations (contact to contact, pooling across the three contact types, and noncontact to noncontact) minus the correlation between contact and noncontact conditions. Higher positive values of this decoding index indicate a stronger distinction between contact and noncontact relationships. In PN, we found significant contact versus noncontact decoding that generalized across naturalistic and rendered scenarios (decoding index: mean ± SEM = 0.045 ± 0.01, $P = 0.0002$ on a two-sided Wilcoxon signed-rank test for zero median; Fig. 2B). This decoding index was positive in 13 of the 14 participants. Although contact versus noncontact decoding was significant in LOC (decoding index: mean ± SEM = 0.022 ± 0.007, $P = 0.0085$ for a two-sided Wilcoxon signed-rank test), it was not so in VTC (mean ± SEM = 0.016 ± 0.007, $P = 0.24$ for a two-sided Wilcoxon signed-rank test in VTC; Fig. 2B). To test whether the decoding found in PN was significantly greater than other fROIs tested, we conducted an analysis of variance (ANOVA) on the decoding indices computed across real-world and rendered

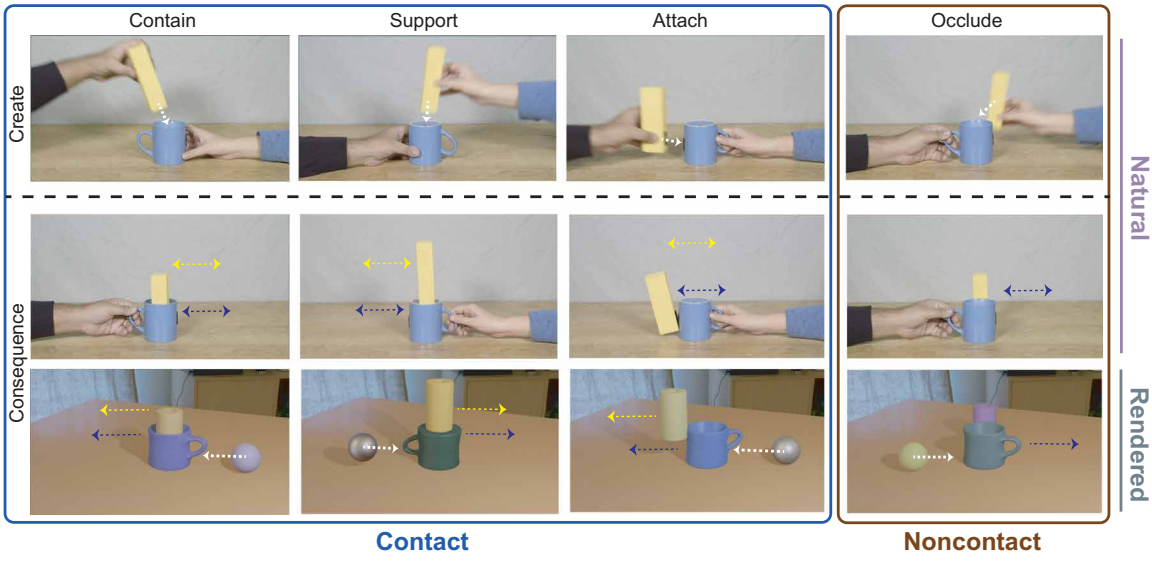

Fig. 1. Stimuli used in experiment 1. Frames from example video clips showing the creation (top row) or consequences (bottom two rows) of contact (containment, support, or attachment) or noncontact (occlusion) relations between objects. Arrows depict the motion trajectories of the objects.

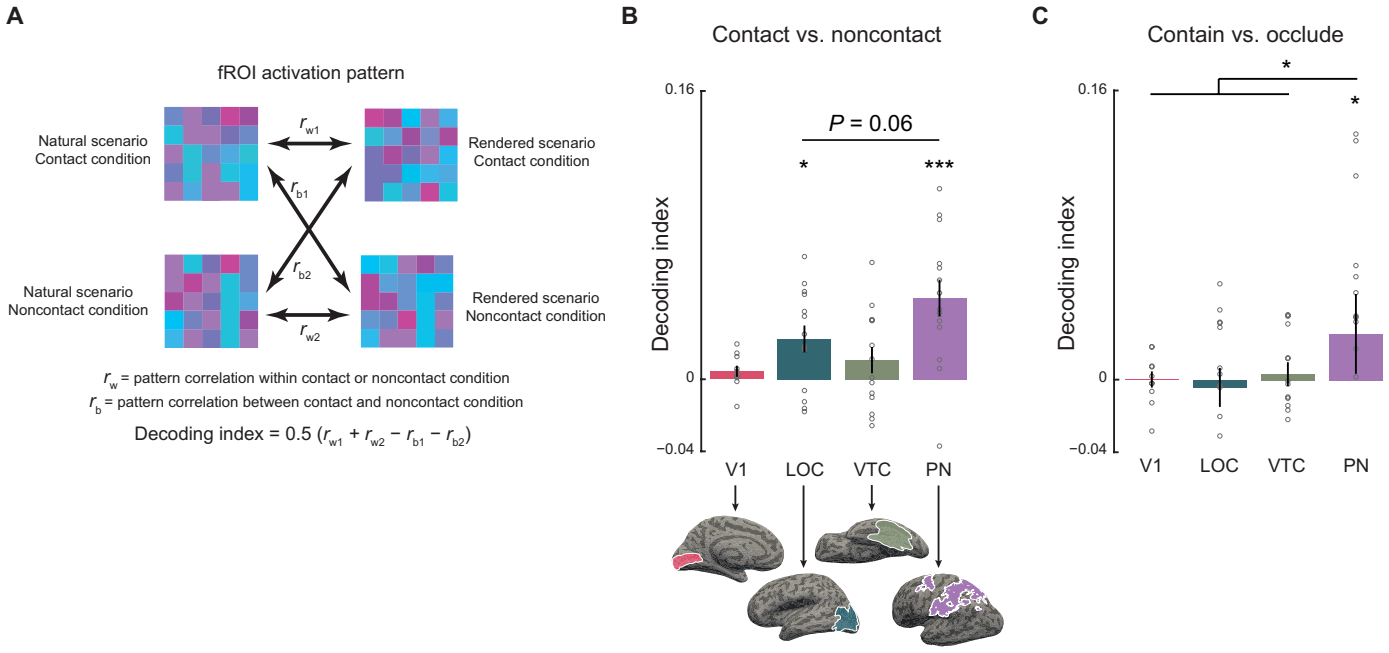

Fig. 2. Contact decoding in experiment 1. (**A**) Our scenario-invariant decoding index is calculated for a given fROI as the correlation in the pattern of response within contact relationships (contact to contact and noncontact to noncontact) minus the correlation between contact relationships (contact to noncontact) when these correlations are computed across scenarios (natural to rendered). (**B**) Contact information by this measure is significant in LOC and PN, but not in V1 or VTC, and is significantly greater in PN than in all three other fROIs. The group parcel for each ROI is shown below on the inflated brain surface of the left hemisphere. (**C**) When the decoding index is computed for just the most visually similar pairs (contain versus occlude), contact information remains significant in PN but not in any of the other fROIs. In (B) and (C), circles represent individual participants; $*P < 0.05$ and $***P < 0.0005$.

scenarios with fROI as the factor. We found a significant effect of fROI [$F(3) = 4.01$, $P = 0.0085$], indicating that decoding differs across fROIs. Post hoc analysis on the ANOVA model revealed that contact versus noncontact decoding was significantly stronger in PN compared with both VTC and V1 (primary visual cortex) ($P = 0.005$ and 0.004, respectively) and marginally significant compared with LOC ($P = 0.06$). Furthermore, the decoding index did not reach significance in V1 (decoding index: mean ± SEM = 0.0044 ± 0.0031, $P = 0.13$ on a two-sided Wilcoxon signed-rank test for zero median; Fig. 2B), arguing against low-level features as the basis of decoding in PN. Thus, PN carries scenario-invariant information that can distinguish contact from noncontact relationships.

PN is a broad region spanning both hemispheres of frontoparietal cortices. Are the decoding results driven by a specific set of subregions within PN? An ANOVA on decoding indices computed across real-world and rendered scenarios in PN with hemisphere (left and right) and lobes (frontal and parietal) as factors revealed neither of the main effects nor the interaction of these two factors ($P > 0.1$). Thus, we do not detect subregional differences in decoding performance.

Univariate responses to contact and noncontact conditions. Are the MVPA results driven by stronger responses to one condition (e.g., contact) over the other (e.g., noncontact)? To answer this question, we computed the average activations within each fROI for contact and noncontact conditions separately. We found that contact and noncontact conditions were not significantly different from each other in any of the four fROIs tested ($P > 0.1$, for a pairwise t test on average activations within an fROI to contact and noncontact conditions across participants).

Contain versus occlude decoding

Contact decoding was significant not only in PN but also in LOC—a region in the ventral visual pathway that is known to represent object shape (*14*). Could the significant contact decoding observed in LOC (and hence, PN) be driven by the differences in composite shapes formed by the two objects in contact and noncontact relationships rather than actual contact itself? For instance, the composite formed by the two objects in support and attach relationships is taller and wider, respectively, compared with the composite entity in the occlude relationship. To reduce the effect of shape on contact decoding, we restricted our analysis to only the contain and occlude relationships that minimally vary in shape but crucially, for our purpose, vary in contact. As before, we used the correlation-based MVPA (see Fig. 2A) to derive our decoding measure, but instead of all three contact relationships, we used responses only for the contain relationship. Here, too, we found significant contact decoding in PN (decoding index: mean ± SEM = 0.026 ± 0.022, $P = 0.016$ on a two-sided Wilcoxon signed-rank test for zero median; Fig. 2C) but not in the other fROIs tested. Furthermore, the decoding index was significantly higher in PN compared with other fROIs ($P < 0.05$, Wilcoxon rank sum test for equal medians for PN with every other fROI tested). Thus, scenario-invariant contact decoding is unlikely to be driven by shape-related differences in PN.

Furthermore, both contact versus noncontact and contain versus occlude decoding was robust to the definition of the fROI. We found significant decoding in PN when restricting our analysis to either the top 10% or top 100 voxels based on the strength of the localizer contrast within each fROI (fig. S1).

Decoding contact type

Our main preregistered analysis described above demonstrates information about the presence (versus absence) of contact in PN. Does PN also distinguish one contact type from another? We answered this question by computing the decoding index (Fig. 2A) for every pair of contact relationships (contain versus support, contain versus attach, and support versus attach). This contact-type decoding—averaged across all three pairs of contact relationships—did not reach significance in either PN (decoding index: mean ± SEM = 0.0086 ± 0.013, $P = 0.81$ on a two-sided Wilcoxon signed-rank test for zero median; fig. S2) or VTC (decoding index: mean ± SEM = 0.0198 ± 0.012, $P = 0.1$ on a two-sided Wilcoxon signed-rank test for zero median) but was significant in LOC (decoding index: mean ± SEM = 0.025 ± 0.0084, $P = 0.0023$ in a two-sided Wilcoxon signed-rank test for zero median). This decoding in LOC was not significantly higher than in PN ($P = 0.22$, in a two-sided Wilcoxon signed-rank test). However, we also found significant contact-type decoding in V1 (decoding index: mean ± SEM = 0.015 ± 0.0068, $P = 0.037$ on a two-sided Wilcoxon signed-rank test for zero median), indicating that the decoding effect in LOC could be driven by differences in low-level visual features (although our blocked design was neither ideal nor intended for a linear Support Vector Machine (SVM) decoding analysis, which was not preregistered, we conducted this analysis in response to a reviewer and found similar but statistically weaker contact-type decoding results). The absence of information in PN (but its presence in V1 and LOC) about the specific kind of contact relationship could indicate that PN abstracts away from both object shape and contact type while maintaining information only about either the presence or absence of contact. Although these results could reflect an important representational difference between LOC and PN, caution is warranted in interpreting these results because contact-type decoding is also significant in V1, suggesting that decoding in LOC could reflect low-level visual confounds, and because contact-type decoding was not significantly greater in LOC than in PN.

Experiment 2: Prediction of future contact in PN

In the second experiment, we sought to test the hypothesis that PN is engaged in forward simulation. To do this, we constructed new video stimuli in which contact events (i.e., collisions) in two different scenarios (roll and throw) were either explicitly shown or not shown but predicted to happen next (see Fig. 3). We then collected fMRI responses for each voxel in each fROI to contact and noncontact events across conditions (perceived and predicted) and scenarios (roll and throw). We then used these responses to test whether representation of a predicted contact event is similar to that of an actually perceived contact event, as predicted by the hypothesis that PN is a physics engine that simulates what will happen next.

Decoding of future contact events in PN

Within the "roll" or "throw" scenario. As a first test of our hypothesis that PN predicts future events, we tested whether predicted contact versus noncontact events can be decoded from perceived contact versus noncontact events within a scenario (either roll or throw) in PN. Specifically, we used a correlation-based MVPA to probe contact decoding that generalized across perceived and predicted conditions within either roll or throw scenario (Fig. 4A; see Materials and Methods). The rationale here is that a brain region involved in forward simulation should evoke similar patterns of activations for both seen and predicted events. In PN, we indeed found significant decoding of contact versus noncontact events across perceived and predicted conditions (decoding index: mean ± SEM = 0.051 ± 0.017, $P = 0.013$ on a two-sided Wilcoxon signed-rank test for zero median; Fig. 4B). In contrast, future contact decoding was not significant in the ventral visual fROIs (decoding index: mean ± SEM = −0.007 ± 0.015, $P = 0.9$ for a two-sided Wilcoxon signed-rank test in LOC and mean ± SEM = −0.018 ± 0.01, $P = 0.11$ for a two-sided Wilcoxon signed-rank test in VTC; Fig. 4B). The decoding index was also not significant in V1 (decoding index: mean ± SEM = 0.0147 ± 0.0136, $P = 0.38$ on a two-sided Wilcoxon signed-rank test for zero median; Fig. 4B), indicating that future contact decoding in PN is unlikely due to low-level visual features. Moreover, the decoding index was significantly higher in PN than in V1, LOC, and VTC ($P < 0.05$, Wilcoxon rank sum test for equal medians for each pairwise comparison of fROIs). This dependence of decoding on fROI was further supported by an ANOVA on decoding indices across participants with fROI (V1, LOC, VTC, and PN) as the factor, which revealed a significant effect of fROI [$F(3) = 6.86$, $P = 0.0002$]. Thus, PN, and not other visual regions, carries information about future contact versus noncontact

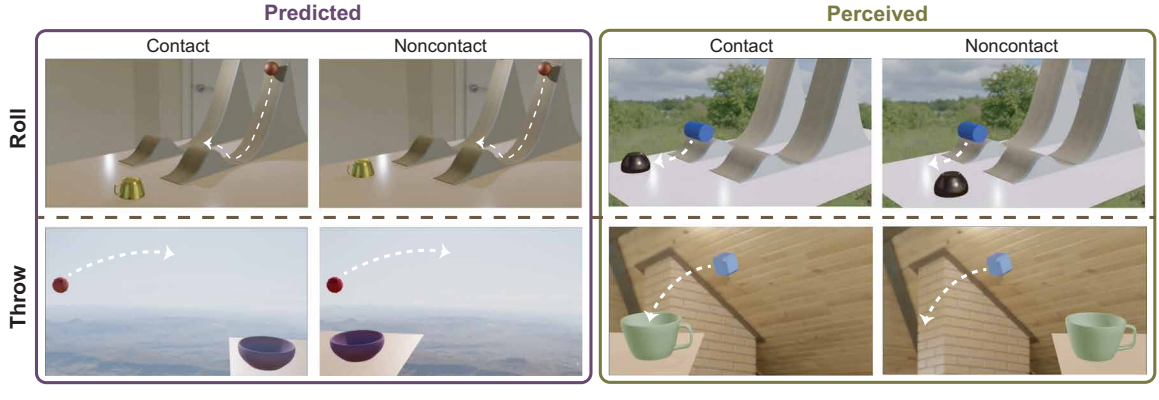

Fig. 3. Stimuli used in experiment 2. Frames from example video clips showing contact versus noncontact events explicitly (perceived), or showing events from which contact or noncontact is predicted to happen next (predicted), for two scenarios, roll and throw. The logic of this experiment is that if PN represents a future predicted contact event, then we should find that the pattern of response in PN for predicted contact stimuli should resemble the pattern of response when a contact event is actually perceived. Arrows indicate the motion trajectory of the agent object. Under the predicted condition, the arrowheads indicate the final position of the object before the video cuts off.

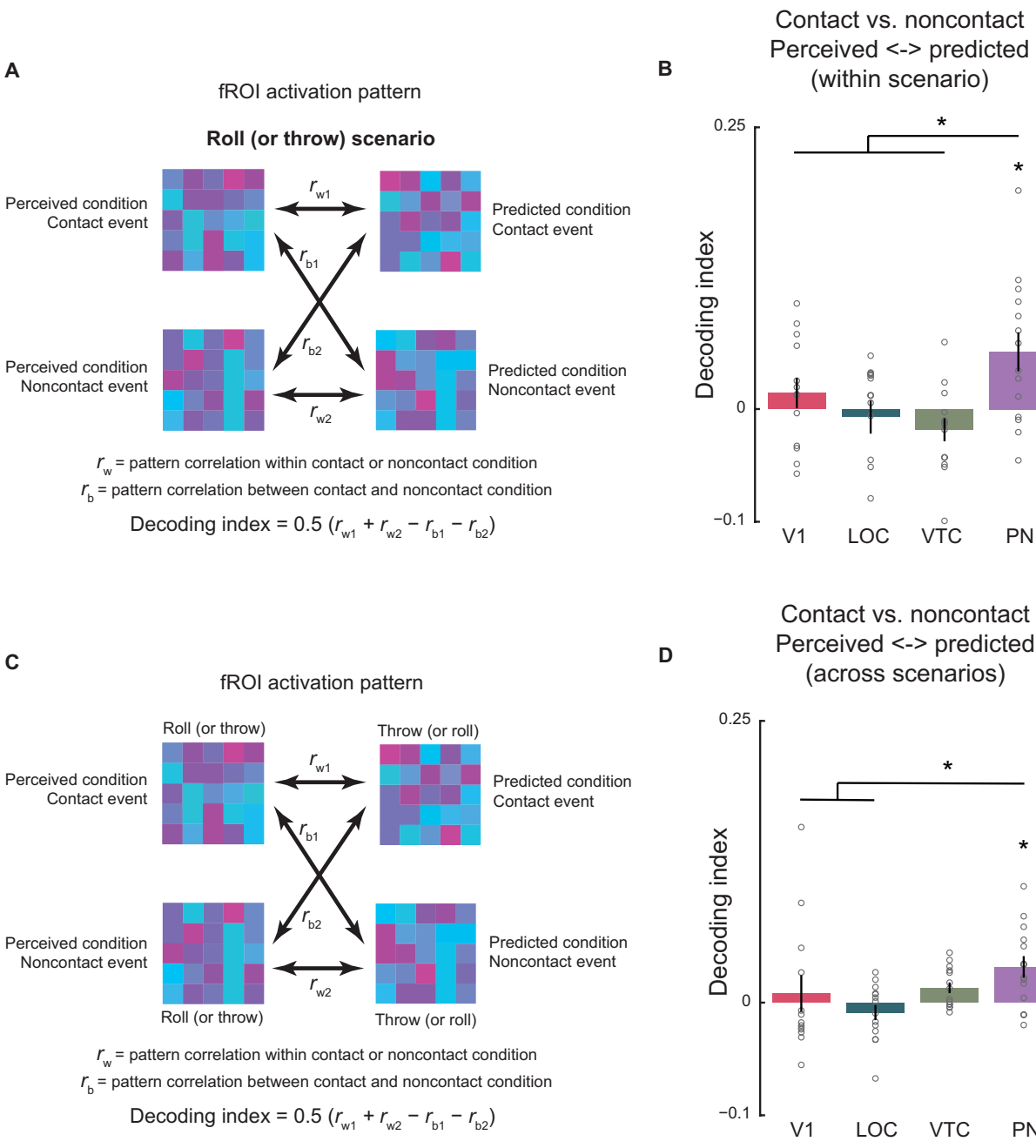

Fig. 4. Contact decoding and prediction in experiment 2. (**A**) Our scenario-invariant decoding index is calculated for a given fROI as the correlation in the pattern of response within contact relationships (contact to contact and noncontact to noncontact) minus between contact relationships (contact to noncontact) when these correlations are computed across perceived and predicted conditions. (**B**) Contact information as measured by the decoding index averaged over both roll and throw scenarios is significant in PN, but not in V1, LOC, or VTC, and is significantly greater in PN than in all three other fROIs. (**C**) Similar schematic as in (A) but the correlations are now computed across both conditions (predicted and perceived) and scenarios (roll and throw). (**D**) Contact information is significant only in PN. *$P < 0.05$.

events with pattern responses like those during perceived contact versus noncontact events within each scenario. Of course, it is always possible that the prediction of some other physical state (other than contact or collision) that we have not tested here may turn out to be made in LOC and other visual regions.

Across roll and throw scenarios. A skeptic could explain away the result just described as a consequence of trivial extrapolation—the agent object continues along the same trajectory under the predicted condition as it did under the perceived condition. We addressed this issue by asking whether future contact decoding generalizes from perceived to predicted conditions not only within a scenario (roll or throw) but also across scenarios (roll to throw or vice versa) wherein the agent objects move along different trajectories. As before, we used a correlation-based MVPA decoding approach (Fig. 4C; see the Materials and Methods) with a positive decoding index within an fROI indicating similar voxel pattern activations for contact (or noncontact) events

across not only predicted to perceived conditions but also roll to throw scenarios. Supporting our hypothesis, we found significant generalizable decoding of future contact events only in PN (decoding index: mean ± SEM = 0.032 ± 0.01, $P = 0.0052$ on a two-sided Wilcoxon signed-rank test for zero median; Fig. 4D) and not in the ventral visual fROIs (decoding index: mean ± SEM = −0.009 ± 0.007, $P = 0.33$ for a two-sided Wilcoxon signed-rank test in LOC and mean ± SEM = −0.013 ± 0.005, $P = 0.07$ for a two-sided Wilcoxon signed-rank test in VTC; Fig. 4D). Cross-scenario decoding of perceived to predicted contact versus no-contact events was also not found in V1 (decoding index: mean ± SEM = 0.008 ± 0.016, $P = 0.9$ on a two-sided Wilcoxon signed-rank test for zero median; Fig. 4D), arguing against an account of our effect in terms of low-level visual confounds. Furthermore, the decoding index was significantly higher in PN than in V1 and LOC ($P < 0.05$, Wilcoxon rank sum test for equal medians) and was only marginally higher compared with VTC ($P = 0.09$, Wilcoxon rank sum test for equal medians). This dependence of decoding on fROI was further supported by an ANOVA on decoding indices computed across roll and throw scenarios with fROI (V1, LOC, VTC, and PN) as a factor, which revealed a significant effect of fROI [$F(3) = 3.71$, $P = 0.0125$]. As in experiment 1, an ANOVA on decoding indices across participants with hemisphere (left versus right) and lobe (frontal versus parietal) as factors found no significant main effects or interactions ($P > 0.1$). Furthermore, a searchlight analysis also revealed decoding primarily in the frontoparietal cortices—regions highly overlapping with PN (see the "Searchlight analysis" section in the Supplementary Materials). Thus, our results indicate that PN carries information about future contact events in an abstract manner that generalizes across scenarios (which is not mere extrapolation of motion trajectories), providing further evidence that these brain regions are involved in predicting future states of the world through forward simulation.

Decoding contact from noncontact events across roll and throw scenarios included perceived and predicted conditions with both leftward and rightward trajectories of objects. We conducted a stronger test of generalization by decoding not only across conditions and scenarios but also object motion trajectories. That is, we used the correlation-based MVPA decoding approach as before, computing pattern correlations for contact (or noncontact) events across perceived and predicted conditions, roll and throw scenarios, and, crucially, across leftward and rightward motion trajectories. Here, too, we found significant decoding of future contact events generalizable across conditions, scenarios, and motion direction only in PN (decoding index: mean ± SEM = 0.033 ± 0.02, $P = 0.008$ on a two-sided Wilcoxon signed-rank test for zero median; fig. S3) and in neither the ventral visual fROIs (decoding index: mean ± SEM = −0.01 ± 0.01, $P = 0.54$ for a two-sided Wilcoxon signed-rank test in LOC and mean ± SEM = −0.002 ± 0.007, $P = 0.64$ for a two-sided Wilcoxon signed-rank test in VTC) nor V1 (decoding index: mean ± SEM = 0.0003 ± 0.012, $P = 0.52$ on a two-sided Wilcoxon signed-rank test for zero median). The decoding index in PN was significantly higher than in other fROIs tested ($P < 0.05$). Thus, the abstract information about future contact events in PN generalizes across object motion direction.

Univariate responses to contact and noncontact events. To test whether the MVPA results were driven, in part, by univariate differences between contact and noncontact events, we computed the average activations within each fROI for contact and noncontact events separately. We found that contact and noncontact events under both perceived and predicted conditions were not significantly different from each other in any fROI tested (all P values > 0.1, for pairwise t tests on average activations within each fROI to contact and noncontact events, either under the perceived or predicted condition, across participants), except in V1 and LOC for the perceived condition ($P < 0.05$, for a paired t test comparing the average responses to contact and noncontact events across participants; noncontact > contact in V1 and contact > noncontact in LOC). Thus, both contact and noncontact events elicited similar responses in PN.

Contact information in PN using representational similarity analysis

As a further test of abstract contact decoding, we used representational similarity analysis (RSA) (*15*) to compare the overall representational structure in various brain regions to an ideal model that perfectly distinguishes between contact and noncontact events across scenarios (roll and throw) and conditions (perceived and predicted). The ideal contact representational dissimilarity matrix (IC-RDM; see Materials and Methods for details on how it was computed) showed a significant correlation with fMRI RDM only in PN (average RDM correlation across participants = 0.023, $P < 0.05$), and this correlation was significantly higher than in LOC ($P < 0.05$ on a paired t test on RDM correlations across participants). Thus, a representational similarity analysis also provides evidence for abstract contact information in PN.

How much of this abstract contact information can be explained by visual features alone? To find out, we computed an RDM containing contact discriminability based on features from a video foundation model RDM (VM-RDM; see the Materials and Methods for details) and compared this with RDM computed on fMRI pattern activations in each fROI (fMRI RDM). We found that VM-RDM was marginally significantly correlated with fMRI RDM only in PN (average RDM correlation across participants = 0.012, $P = 0.06$), indicating that predicted contact decoding in PN may be at least partially driven by visual perceptual features. However, the correlation between the fMRI RDM in PN and the IC-RDM remained significant even after partialling out VM-RDM (average correlation across participants = 0.022, $P < 0.05$), indicating that PN represents and predicts contact information over and above those captured by visual perceptual features.

DISCUSSION

In this study, we used object contact relationships to test whether the hypothesized PN runs forward simulations to predict what will happen next, a core tenet of the hypothesis that PN constitutes the brain's physics engine. In experiment 1, we showed that PN carries abstract information about object contact—a critical attribute for forward simulation. In experiment 2, we showed that PN not only carries information about perceived contact events (replicating and generalizing our results from experiment 1) but also shows similar patterns of response for contact events that are merely predicted but not seen. In both experiments, our decoding results (i) generalized across objects and scenarios, (ii) were obtained even though participants were performing an orthogonal (one-back) task, and (iii) were weaker or absent in the ventral visual pathway (LOC and VTC). Together, our findings demonstrate that PN encodes abstract object contact information, and provide the strongest evidence to date that PN runs forward simulations to predict what will happen next.

Several lines of evidence indicate that our findings do not simply reflect visual feature confounds. First, in both experiments, our decoding results were not significant in V1. Second, in experiment 2, our main predicted contact decoding results were computed across roll and throw scenarios (Fig. 4D), where relative spatial positions and motion trajectories of objects were unconfounded from contact and noncontact relationships. Third, we show significant predicted contact decoding in PN that generalizes across both object trajectories (leftward versus rightward) and scenarios (roll versus throw). Thus, our decoding results in PN point to an abstract representation of object contact, not low-level visual confounds.

Representation of object relations in the mind and brain

Our evidence for contact representation in PN dovetails with substantial literature on the primacy of physical object contact relationships in language and in visual scene understanding across the life span. Infants (~3 to 6 months) are sensitive to containment (*7*) and support (*16*, *17*) relationships. Behavioral studies [see (*6*) for a review] of object relations have shown that adults (i) recognize them from extremely brief exposures (*18*), (ii) encode them in a categorical manner over and above equivalent metric changes in the stimuli (*19*, *20*), and (iii) alter their attentional deployment during visual search to follow object relations (*21*). One study (*22*) found particularly compelling evidence that object relations are coded automatically and abstractly: Participants were asked to identify a target image (e.g., of a phone inside a basket) among a stream of distractors false alarmed to images containing the same object relationship, although the objects were completely different (e.g., a knife sitting inside a cup). Despite this wealth of behavioral evidence on the importance of object relational attributes in perception, very few studies have investigated their neural representation. One fMRI study showed that attending to categorical spatial relationships (like above/below and behind/in front of) compared with attending to the identity of objects resulted in increased activity in the left parietal cortex and bilateral posterior middle frontal cortex (*23*). Other fMRI studies have shown that parts of the ventral visual pathway are sensitive to relative positions of two objects on the screen (*24*) and familiar configuration of objects (*25*). Another recent fMRI study has shown that large parts of the ventral visual pathway (including V1 and LOC) and parietal cortices have information about events involving two objects (*26*). A few neuropsychological studies have also shown that damage to the left inferior parietal and prefrontal cortices can lead to deficits in processing visual spatial relations and locative prepositions (*23*, *27*). Our results build upon this work to show that both LOC and PN carry information about object contact in a manner that is generalizable across objects and scenarios and even when the task does not explicitly require it.

Physical prediction

Our most important finding is that PN represents predicted contact (versus noncontact) similarly to the way it represents perceived contact (versus noncontact). This finding goes beyond showing that PN is merely engaged during a physical prediction task (as in the case of the "intuitive physics" localizer), by demonstrating that this network contains an explicit representation of the content of the prediction, i.e., of what will happen next. The fact that we see abstract information about predicted contact events in PN and not the ventral visual pathway suggests that the computations carried out in the latter are unlikely to be the sole neural substrate for generalizable physical reasoning. Thus, our findings provide the strongest evidence to date that PN might contain a generative physical model of the world that runs forward simulations. These results further invite multiple new lines of inquiry.

First, if PN is indeed involved in building a generative model of the world and running forward simulations to predict what will happen next, then how abstract are the representations in this internal model of the world? It has been hypothesized that the mental physics engine intelligently compresses the rich details of the physical world to a small set of entities and events to efficiently process information for human perception and generate suitable predictions for action planning and intervention (*11*). These approximations and abstractions can also enable faster predictions through parallel simulations on a small number of tokens, leading to our ability to make rapid and automatic physical inferences (*18*, *28*). Work in robotics has used similarly abstract representations of object relations for planning (*29*). Our finding that both perceived and predicted contact could be decoded across objects and scenarios indicates that this abstract information is present in PN, as predicted by the game engine hypothesis.

Note, however, that these abstract representations are not sufficient for a forward simulation, which would require additional precise representations of specific types of contact, mass, trajectories, forces, etc. Given the null result in PN for specific contact-type decoding, we see at least two possible ways of interpreting our current results in the context of what different levels of abstraction might mean for future prediction. One possibility is that the latents of the physical world are represented in a hierarchical manner where both abstract and specific information are represented simultaneously, to be read off for efficient processing depending on task demands. These hierarchical representations could be present either only within PN or in PN along with other parts of the brain (like the ventral visual pathway). Our results seem most consistent with the latter, where PN has abstract information about contact/noncontact relationships and LOC has information about the specific contact types. Another possibility is that representations in PN are modulated by task demands, and if the specific contact type was necessary for subsequent prediction or downstream planning tasks, then we would find that information in PN. Our current experiments do not distinguish between these possibilities. However, future research can test whether and how PN integrates abstract information with more detailed and quantitative information as needed for rich forward simulations.

Second, how far into the future are predictions made? Intuition suggests that we do not make multistage predictions routinely or automatically, although planning multistep complex routines in the real world may often require it, such as when planning a bank shot in pool or when making eggs for breakfast. However, it may be that in everyday perception and planning we predict in hierarchical stages: using simulation to predict the state of the world up until an event boundary [for example, marked by a contact event; see (*30*)] and then organizing these local temporal predictions into a higher-level multistage symbolic routine for action planning.

Third, how automatic are these predictions? A recent study found that blood oxygen level–dependent (BOLD) responses in motion-selective areas in the human brain when participants simulate a ball's trajectory are similar to responses when participants actually perceive the ball's trajectory (*31*). However, whereas participants in that study were actively engaged in a simulation-based task, in our study, participants were not asked explicitly to predict what

would happen next. Our participants were simply asked to look at the stimuli and remember them for a few seconds to perform a one-back task. Our study provides the first piece of evidence that similar brain regions are involved in predicting future events even during an orthogonal task that did not require active simulation. Future studies can test whether predictions are made even for unattended stimuli, for multiple objects simultaneously, or in the face of competing task demands.

Fourth, neither fMRI nor electrophysiological recordings in animals can answer the important question of the causal role of PN in the representation of object relations and physical prediction. Studies of neurological patients with damage in these regions will be important to address this question in the future.

Perhaps the most important open question is how exactly we make physical predictions. In an energetic ongoing debate, some cognitive scientists have argued that physical prediction does not entail rich forward simulation but instead engages a simpler and faster mechanism akin to pattern recognition (*18*, *32*–*36*). The fact that we did not find evidence of prediction in the ventral pathway (i.e., LOC and VTC), which is most clearly implicated in pattern classification (*37*, *38*), is suggestive that the predictive information we find in the dorsal pathway is based on a different computational strategy. It will be important in the future to test neural responses in PN against computational models of physical simulation. A study in macaques took a first step in this direction and found that neural responses recorded from macaque dorsomedial frontal cortex (DMFC) during a Mental-Pong game better match representational dynamics in recurrent neural networks (RNNs) that were explicitly trained to approximate simulations of the occluded ball's intermediate positions, compared with RNNs that were trained only to predict the ball's end point (*39*, *40*). Further evidence suggests that video foundation models that leverage the temporal structure in naturalistic tasks have latent representations that better capture neural dynamics in macaque DMFC during the Pong task (*41*). Given that the neural evidence thus far is suggestive of both abstraction and precision in forward prediction, computational models that aggregate results over multiple timescales and levels of abstraction could be required to explain physical prediction in the brain.

Relation of PN to other brain regions and networks

Concerning PN more broadly, it will be important to better understand its relationship to other brain networks involved in action planning, tool use, and processing cognitively demanding tasks [i.e., the multiple demand (MD) system]. We find that the group-level activation maps for physical inference (physics > color task in our localizer) overlap with regions previously shown to be engaged during visually guided action (*42*), tool use (*43*–*45*), and other broad range of cognitive tasks (*46*, *47*). It is plausible that visually guided action, tool use, and physical reasoning share the same computational goals and constraints (*48*). However, the apparent overlap of these regions with the "MD" network also raises the possibility that physical inference (and action planning and tool use) could be recruiting general-purpose computational and cognitive machinery. However, another possibility is that instead of being a homogeneous network, the MD system is further fractionated into systems with distinct subregions engaged in physical reasoning, action planning, etc. Although preliminary behavioral evidence suggests that mental faculties involved in physical reasoning and working memory are distinct (*49*), future studies can explore how this result relates to neural representations.

In conclusion, our study not only shows that PN contains abstract information about object contact but also provides the strongest evidence yet that these brain regions are engaged in predicting what will happen next. Together, these findings support and enrich the hypothesis that these brain regions serve as a physics engine in the brain, supporting our ability to understand and predict the world around us.

MATERIALS AND METHODS

fMRI data acquisition

All imaging was performed on a Siemens 3T Prisma scanner with a 32-channel head coil at the Athinoula A. Martinos Imaging Center at the Massachusetts Institute of Technology (MIT). For each participant, a high-resolution T1-weighted anatomical image [Magnetization Prepared RApid Gradient-Echo (MPRAGE): repetition time (TR) = 2.53 s, echo time (TE) = 3.57 ms, a = 9°, field of view (FOV) = 256 mm, matrix = 256 × 256, slice thickness = 1 mm, 176 slices, acceleration factor = 2, 24 reference lines, bandwidth (BW) = 190 Hz per pixel] was collected in addition to whole-brain functional data using a T2*-weighted echo planar imaging pulse sequence (TR = 2 s, TE = 30 ms, a = 90°, FOV = 204 mm, matrix = 102 × 102, slice thickness = 2 mm, voxel size = 2 mm by 2 mm in plane, slice gap = 0 mm, 66 slices).

fMRI data preprocessing

All basic preprocessing steps and general linear model (GLM) analyses were similar to our previous work (*3*). In experiment 1, in addition to the run-wise and motion nuisance regressors, the GLM included regressors for each of the 27 experimental conditions. In experiment 2, we fit separate GLMs for runs containing perceived and predicted conditions (see the corresponding "Stimuli and experimental design" section). Specifically, separately for each run type (i.e., perceived or predicted), the GLM included individual regressors for all 48 stimuli with an additional regressor modeling the BOLD activation during the one-back task trials. These 49 regressors were considered in addition to the standard run-wise and motion nuisance regressors. All other analyses were performed in MATLAB 2018b (MathWorks).

fROI definition

All ROIs were localized independently from the main experiment. The hypothesized PN was functionally defined using two runs of an intuitive physics fMRI localizer task (*1*). We identified PN in each participant individually using the physics task > color task contrast of the localizer (uncorrected $P < 0.001$). We then intersected the significance map with group-level parcels created from the localizer data in previous studies (*2*, *3*). We analyzed PN as a whole and looked for effects in its subregions (frontal and parietal parts) separately. LOC was functionally defined using a dynamic face, object, scenes, and scrambled objects (dynFOSS) localizer (*50*). Specifically, we identified LOC in each participant individually using the objects > scrambled objects contrast (uncorrected $P < 0.001$). We then intersected this significance map with masks derived from anatomical parcellation. We analyzed responses across both hemispheres. VTC was also functionally defined by intersecting the significant voxels in the all visual > fixation contrast (uncorrected $P < 0.001$) from the dynFOSS localizer and anatomical parcellation from the

"Desikan-Killiany" atlas in FreeSurfer. We analyzed responses across both hemispheres. The average number of voxels in each of these functionally localized ROIs is shown in Table 1. V1 was also defined using the dynFOSS localizer. We identified V1 by intersecting anatomically derived parcels (using FreeSurfer labels) with the scrambled > objects contrast (uncorrected $P < 0.001$). We analyzed responses across both hemispheres.

We also repeated our main analyses by taking either the top 10% or top 100 voxels within each parcel based on the *t* values of relevant contrast (see fig. S1). This analysis controlled for the number of voxels in each fROI and participant.

Contact decoding experiment (experiment 1)

The stimulus design and analysis methods were preregistered (https://osf.io/ezq3s) before running the full experiment.

Participants

Fourteen participants (ages 22 to 38 years; 8 females) participated in the fMRI experiment. The sample size was determined on the basis of a power analysis of contact versus noncontact decoding within the hypothesized PN in the pilot data ($n = 5$). The power analysis was performed using G*Power (www.psychologie.hhu.de/arbeitsgruppen/allgemeine-psychologie-und-arbeitspsychologie/gpower) with a statistical power ($1 - \beta$) of 95% at a significance level (α) of 0.01 to reach the effect size (*d*) of 1.23 observed in our pilot data. All participants had normal or corrected-to-normal vision. Before participating in the experiment, all participants gave informed consent to the experimental protocol approved by the MIT Committee on the Use of Humans as Experimental Subjects (no. 0403000096). The study was conducted in compliance with all the relevant ethical guidelines and regulations for work with human participants. An additional three participants were scanned for this experiment, but their data were excluded from further analyses due to excessive motion in one of the participants and failure to independently localize PN in the other two.

Stimuli and experiment design

Each participant completed one 2-hour scan session, which consisted of the following: (i) a high-resolution anatomical scan, (ii) two runs of the PN localizer (*1*), (iii) one run of the dynFOSS localizer (*50*), and (iv) six to eight runs of the contact decoding experiment. The contact decoding experiment had four main relationship types: containment, support, attachment, and occlusion. Each relationship type was embedded in three different scenarios (natural-create, natural-consequence, and rendered) using two different object types (bowl and mug). Thus, there were 24 unique relationship conditions. We included three additional single-objects–only conditions (bowl, mug, and others) as a baseline for object decoding. The natural-create scenario depicted a hand holding the base object (mug or bowl), while another hand placed a second object in the scene to create the relationship. The natural-consequence scenario depicted a hand moving the base object back and forth (Fig. 1), with the second object already in position to reveal the relationship and physical contingency between the two objects. The rendered scenario revealed the physical contingency using a ball colliding with the base object. The natural scenarios were filmed in-house with human actors, whereas the rendered scenario was created using Blender. Using different exemplars of objects and counterbalancing the motion trajectories of hands and balls from the left and right side of the frame, we created 768 video clips across all conditions each lasting 3 s. Each block of the main experiment consisted of four 3-s videos from 1 of the 27 conditions (24 object-object relationship conditions + 3 single-object conditions). Specifically, we randomly chose four video clips from a subcondition and presented them in a sequence within the block. The video clips within each block were randomized independently for each participant and chosen with replacement across blocks within a participant. One of the video clips was repeated consecutively within each block, and the participants were instructed to press a button whenever they detected such repetitions (i.e., an orthogonal one-back task). Thus, each block lasted for 15.4 s (five 3-s videos with a 100-ms interstimulus interval). Each run consisted of 28 stimulus blocks (the "others" single-object condition was repeated twice within a run to equalize the amount of data for base objects and secondary objects), and 5 fixation blocks (15 s each) interspersed uniformly within the run (one in the beginning and one after every seventh stimulus block). Throughout the run, the participants were instructed to maintain fixation on a red dot at the center of the screen to minimize eye movement confounds. The order of blocks within a run was determined using a Latin square design, which also helped counterbalancing the condition sequence across runs.

Contact decoding

To test whether contact relationships can be distinguished from noncontact relationships within an fROI, we used a correlation-based MVPA (*51*). We computed the response of each voxel within an fROI for contact relationship by averaging its response to all the contact relationships (containment, support, and attachment) across base objects (bowl and mug) within each scenario (natural-create, natural-consequence, and rendered) separately. Similarly, we computed the response of each voxel for the noncontact relationship by averaging its response to the occlusion relationship across the base objects within each scenario separately. We then computed the within (r_{w1} and r_{w2}) and between (r_{b1} and r_{b2}) contact-type correlations across the voxels within the fROI and across the natural-create and rendered scenarios (Fig. 2). We then normalized the correlation values (using Fisher *z*-transform) and computed the difference between the averaged within and the averaged between contact-type correlations for each fROI in each participant as our decoding index. A positive value of the decoding index indicates that contact and noncontact relationships evoke distinctive patterns of activations within the fROI. In a similar manner, we computed the decoding index for the pair of natural-consequence and rendered scenarios. The plots (Fig. 2 and fig. S1) contain the decoding indices averaged over pairs of natural-consequence and rendered scenarios.

Table 1. Number of voxels (mean ± SEM) within each fROI across participants for experiments 1 and 2.

	LOC	VTC	PN
Experiment 1	859.9 ± 123.8	975.4 ± 126.8	999.4 ± 231.6
Experiment 2	878.4 ± 170.3	1373.2 ± 146	603.4 ± 192.5

Future contact prediction experiment (experiment 2)

The stimulus design and analysis methods were preregistered (https://osf.io/kr45p/) before running the full experiment.

Participants

Fourteen participants (ages 20 to 35 years; 5 females) participated in the fMRI experiment. All participants had normal or corrected-to-normal vision. Before participating in the experiment, all participants gave informed consent to the experimental protocol approved by the MIT Committee on the Use of Humans as Experimental Subjects (no. 0403000096). The study was conducted in compliance with all the relevant ethical guidelines and regulations for work with human participants.

Stimuli and experiment design

For this experiment, every participant completed a 2-hour scan that included the following: (i) a high-resolution anatomical scan, (ii) two runs of a PN localizer (*1*), (iii) one run of a dynFOSS localizer (*50*), and (iv) eight runs of the future contact prediction experiment—four each of observed and predicted events. The future contact prediction experiment included 96 stimuli, each being a 1.5-s video clip containing two objects (an agent object and a patient object) and varying orthogonally in a 2 × 2 design whether the event type entailed contact or no contact and whether the event was perceived or predicted. These four critical conditions were each shown with three further orthogonally crossed dimensions included to test the generality of the representations: two scenario types (roll or throw), six background scenes (three indoor and three outdoor scenes), and two motion trajectories of the agent object (leftward or rightward). Furthermore, the roll scenario was rendered with a bowl and a cylinder as the agent and patient objects, respectively, under the "perceived" condition and with a mug and a sphere as the agent and patient objects, respectively, under the "predicted" condition to minimize potential low-level visual feature confounds while decoding events within a scenario. Similarly, the throw scenario was rendered with a different pair of agent and patient objects—a mug and a cube for the perceived condition and a bowl and an icosphere for the predicted condition. The stimulus design is summarized in Table 2, and example frames from stimuli for each condition and scenario combination are shown in Fig. 3. Each video was rendered using the Blender software and clipped to 1.5 s either from the start or from the middle to show predicted and perceived conditions, respectively.

Each run of the main experiment showed videos from either only the perceived or predicted condition in an event-related design, and each run included three repetitions of each of the 48 videos for a total of 144 stimulus trials. Each trial consisted of 1.5 s of the video stimuli and a trailing fixation-only period lasting 0.5, 2.5, or 4.5 s. The trial order and the corresponding fixation-only periods were chosen according to optseq2 (https://surfer.nmr.mgh.harvard.edu/optseq/) (*52*). We included 10 one-back trials randomly interspersed within each run to ensure that participants were paying attention to the stimuli. We also included fixation-only blocks each lasting 15 s in the beginning and end of each run.

Every participant was shown four runs each of the perceived and predicted conditions. We specifically sandwiched the four perceived condition runs between two runs each of the predicted conditions. This ensured that not all the predicted condition runs were primed by the perceived condition runs, while ensuring that not all the perceived condition runs were shown later during the scan session, leading to potentially weaker/noisier signals due to participant fatigue.

Contact decoding across perceived and predicted conditions

Within scenario. To test for contact decoding that is generalizable across conditions (perceived versus predicted), we used a correlation-based MVPA (*51*). We computed the response of each voxel within an fROI to contact relationship under the perceived condition (and, say, the roll event) by averaging its response to all the corresponding 12 videos (see Table 2, left-most cell in the first row). Similarly, we computed the response of each voxel in the fROI to the noncontact relationship under the perceived condition (and, again, the roll event) by averaging its response to all the corresponding 12 videos (Table 2, left-most cell in the second row). We then computed the voxel responses to contact and noncontact relationships separately under the predicted condition (also for the roll event) by averaging responses across the corresponding videos (Table 2, second column cells in the first and second rows, respectively). We then gathered the pattern of averaged responses across voxels within the fROI for the four conditions (perceived and predicted events, contact and noncontact relationships within the roll scenario) and computed the within (r_{w1} and r_{w2}) and between (r_{b1} and r_{b2}) contact-type correlations (Fig. 4A). We Fisher *z*-transformed the correlations and computed the difference between the averaged within and averaged between contact correlations for each fROI in each participant as our decoding index. A positive value of this index implies that the contact relationship can be distinguished from the noncontact relationship within the roll scenario across perceived and predicted conditions. We also computed the decoding index for the throw scenario and averaged the two decoding indices within an fROI for each participant.

Across scenario. We extended the within-scenario decoding analysis to measure contact decoding across both conditions (perceived versus predicted) and scenarios (roll versus throw). We computed the response of each voxel within an fROI to contact relationship under the perceived condition (and, say, the roll event) by averaging its response to all the corresponding 12 videos (see Table 2, left-most cell in the first row). Similarly, we computed the response of each voxel in the fROI to the noncontact relationship under the perceived

Table 2. Summary of stimulus design in experiment 2.

	Perceived		Predicted	
Contact	Roll (bowl/cylinder) (6 background scenes × 2 trajectories = 12 stimuli)	Throw (mug/cube) (6 background scenes × 2 trajectories = 12 stimuli)	Roll (mug/sphere) (6 background scenes × 2 trajectories = 12 stimuli)	Throw (bowl/icosphere) (6 background scenes × 2 trajectories = 12 stimuli)
Noncontact	Roll (bowl/cylinder) (6 background scenes × 2 trajectories = 12 stimuli)	Throw (mug/cube) (6 background scenes × 2 trajectories = 12 stimuli)	Roll (mug/sphere) (6 background scenes × 2 trajectories = 12 stimuli)	Throw (bowl/icosphere) (6 background scenes × 2 trajectories = 12 stimuli)

condition (and, again, the roll event) by averaging its response to all the corresponding 12 videos (Table 2, left-most cell in the second row). We then computed the voxel responses to contact and noncontact relationships separately under the predicted condition (this time, for the throw event) by averaging responses across the corresponding videos (Table 2, right-most cells in the first and second rows, respectively). We then gathered the pattern of averaged activations across voxels within the fROI for the four conditions (perceived roll and predicted throw conditions and contact and noncontact relationships) and computed the within (r_{w1} and r_{w2}) and between (r_{b1} and r_{b2}) contact-type correlations (Fig. 4A). We Fisher z-transformed the correlations and computed the difference between the averaged within and the averaged between contact correlations for each fROI in each participant as our decoding index. A positive value for this index implies that contact and noncontact relationships evoke distinctive patterns of activations within the fROI that are generalizable across both scenarios (roll versus throw) and conditions (predicted versus observed). We then computed the decoding index for the observed-throw versus predicted-roll comparison and averaged the two decoding indices within an fROI in each participant.

Representation similarity analysis

As an additional test of abstract contact decoding and to infer the contribution of visual features, we used RSA (*15*) to compare the overall representational structure in various brain regions to (i) a visual feature-based representation extracted from a video foundation model and (ii) an ideal model that perfectly distinguishes between contact and noncontact events across scenarios (roll and throw) and conditions (perceived and predicted).

Video model RDM

We extracted features for each of the 96 video stimuli (perceived and predicted conditions in experiment 2) from each layer of a video foundation model [pfVC1_CTRNN_physion from (*41*)] trained on rendered videos in the Physion benchmark (*53*). For each layer, we trained a linear SVM classifier (fitclinear function in MATLAB) for contact versus noncontact decoding on the feature representations of either the perceived roll or throw scenario and computed the decoding accuracy on the held-out perceived and predicted scenarios. We then computed the average test decoding accuracy in each layer and chose the model layer that showed the maximum test accuracy. This maximum averaged test accuracy was slightly higher than chance (average accuracy = 59.6%, whereas chance accuracy = 50%), indicating that visual features from the model can partially distinguish between contact and noncontact conditions across perceived and predicted scenarios. To get VM-RDM, we computed the average perpendicular distance from the SVM classifier boundary for each stimulus pair in the test set (*54*, *55*).

Ideal contact RDM

We used ground-truth contact and noncontact labels for all stimuli ($n = 96$) under the perceived and predicted conditions of experiment 2 and computed the dissimilarity for each stimulus pair as the absolute difference between the corresponding labels. That is, pairs of stimuli differing in contact were given a dissimilarity value of 1 and other stimulus pairs were given a dissimilarity value of 0.

fMRI RDM

For each ROI in a participant, we computed the representational dissimilarity of the pattern of response across voxels in the fROI for each of the 96 × 96 stimulus pairs as the correlation distance ($1 - r$, where r is the Pearson correlation between the pattern activations).

Supplementary Materials

This PDF file includes:

Figs. S1 to S6

Supplementary Text

REFERENCES AND NOTES

1. J. Fischer, J. G. Mikhael, J. B. Tenenbaum, N. Kanwisher, Functional neuroanatomy of intuitive physical inference. *Proc. Natl. Acad. Sci. U.S.A.* **113**, E5072–E5081 (2016).
2. S. Schwettmann, J. B. Tenenbaum, N. Kanwisher, Invariant representations of mass in the human brain. *eLife* **8**, e46619 (2019).
3. R. T. Pramod, M. A. Cohen, J. B. Tenenbaum, N. Kanwisher, Invariant representation of physical stability in the human brain. *eLife* **11**, e71736 (2022).
4. S. Liu, K. Lydic, L. Mei, R. Saxe, Violations of physical and psychological expectations in the human adult brain. *Imaging Neurosci.* **2**, 1–25 (2024).
5. S. Ullman, N. Dorfman, D. Harari, A model for discovering 'containment' relations. *Cognition* **183**, 67–81 (2019).
6. A. Hafri, C. Firestone, The perception of relations. *Trends Cogn. Sci.* **25**, 475–492 (2021).
7. S. J. Hespos, R. Baillargeon, Reasoning about containment events in very young infants. *Cognition* **78**, 207–245 (2001).
8. R. Baillargeon, "Infants' understanding of the physical world" in *Advances in Psychological Science. Vol. 2: Biological and Cognitive Aspects*, M. Sabourin, F. Craik, M. Robert, Eds. (Psychology Press, 1998), pp. 503–529.
9. E. S. Spelke, A. Phillips, A. L. Woodward, "Infants' knowledge of object motion and human action" in *Causal Cognition* (Oxford Univ. Press, 1996), pp. 44–78.
10. P. W. Battaglia, J. B. Hamrick, J. B. Tenenbaum, Simulation as an engine of physical scene understanding. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 18327–18332 (2013).
11. T. D. Ullman, E. Spelke, P. Battaglia, J. B. Tenenbaum, Mind games: Game engines as an architecture for intuitive physics. *Trends Cogn. Sci.* **21**, 649–665 (2017).
12. I. Yildirim, J. Wu, N. Kanwisher, J. Tenenbaum, An integrative computational architecture for object-driven cortex. *Curr. Opin. Neurobiol.* **55**, 73–81 (2019).
13. L. P. Kaelbling, T. Lozano-Pérez, Integrated task and motion planning in belief space. *Int. J. Robot. Res.* **32**, 1194–1227 (2013).
14. K. Grill-Spector, Z. Kourtzi, N. Kanwisher, The lateral occipital complex and its role in object recognition. *Vision Res.* **41**, 1409–1422 (2001).
15. N. Kriegeskorte, M. Mur, P. Bandettini, Representational similarity analysis – Connecting the branches of systems neuroscience. *Front. Syst. Neurosci.* **2**, 4 (2008).
16. R. Baillargeon, A. Needham, J. Devos, The development of young infants' intuitions about support. *Early Dev. Parent.* **1**, 69–78 (1992).
17. A. Needham, R. Baillargeon, Intuitions about support in 4.5-month-old infants. *Cognition* **47**, 121–148 (1993).
18. C. Firestone, B. J. Scholl, Seeing physics in the blink of an eye. *J. Vis.* **17**, 203 (2017).
19. A. Lovett, S. L. Franconeri, Topological relations between objects are categorically coded. *Psychol. Sci.* **28**, 1408–1418 (2017).
20. J. G. Kim, I. Biederman, C.-H. Juan, The benefit of object interactions arises in the lateral occipital cortex independent of attentional modulation from the intraparietal sulcus: A transcranial magnetic stimulation study. *J. Neurosci.* **31**, 8320–8324 (2011).
21. Y.-H. Yang, J. M. Wolfe, Is apparent instability a guiding feature in visual search? *Vis. Cogn.* **28**, 218–238 (2020).
22. A. Hafri, M. F. Bonner, B. Landau, C. Firestone, A phone in a basket looks like a knife in a cup: Role-filler independence in visual processing. *Open Mind* **8**, 766–794 (2024).
23. P. X. Amorapanth, P. Widick, A. Chatterjee, The neural basis for spatial relations. *J. Cogn. Neurosci.* **22**, 1739–1753 (2010).
24. K. J. Hayworth, M. D. Lescroart, I. Biederman, Neural encoding of relative position. *J. Exp. Psychol. Hum. Percept. Perform.* **37**, 1032–1050 (2011).
25. D. Kaiser, M. V. Peelen, Transformation from independent to integrative coding of multi-object arrangements in human visual cortex. *Neuroimage* **169**, 334–341 (2018).
26. S. Karakose-Akbiyik, A. Caramazza, M. F. Wurm, A shared neural code for the physics of actions and object events. *Nat. Commun.* **14**, 3316 (2023).
27. D. Tranel, D. Kemmerer, Neuroanatomical corellates of locative prepositions. *Cogn. Neuropsychol.* **21**, 719–749 (2004).
28. C. Firestone, B. J. Scholl, Seeing stability: Intuitive physics automatically guides selective attention. *J. Vis.* **16**, 689 (2016).
29. G. Konidaris, L. P. Kaelbling, T. Lozano-Perez, From skills to symbols: Learning symbolic representations for abstract high-level planning. *J Artif. Intell. Res.* **61**, 215–289 (2018).
30. T. S. Yates, S. Yasuda, I. Yildirim, Temporal segmentation and "look ahead" simulation: Physical events structure visual perception of intuitive physics. *J. Exp. Psychol. Hum. Percept. Perform.* **50**, 859–874 (2024).
31. A. Ahuja, T. M. Desrochers, D. L. Sheinberg, A role for visual areas in physics simulations. *Cogn. Neuropsychol.* **38**, 425–439 (2021).

32. E. Ludwin-Peery, N. R. Bramley, E. Davis, T. M. Gureckis, Broken physics: A conjunction-fallacy effect in intuitive physical reasoning. *Psychol. Sci.* **31**, 1602–1611 (2020).
33. E. Davis, G. Marcus, N. Frazier-Logue, Commonsense reasoning about containers using radically incomplete information. *Artif. Intell.* **248**, 46–84 (2017).
34. N. Chater, M. Oaksford, Theories or fragments? *Behav. Brain Sci.* **40**, e258 (2017).
35. E. Ludwin-Peery, N. R. Bramley, E. Davis, T. M. Gureckis, Limits on the use of simulation in physical reasoning, *Proceedings of the Annual Meeting of the Cognitive Science Society* (2019), vol. 41.
36. E. Davis, G. Marcus, The scope and limits of simulation in cognitive models. arXiv:1506.04956 [cs.AI] (2015).
37. M. A. Goodale, A. D. Milner, L. S. Jacobson, D. P. Carey, A neurological dissociation between perceiving objects and grasping them. *Nature* **349**, 154–156 (1991).
38. D. L. K. Yamins, J. J. DiCarlo, Using goal-driven deep learning models to understand sensory cortex. *Nat. Neurosci.* **19**, 356–365 (2016).
39. R. Rajalingham, H. Sohn, M. Jazayeri, Dynamic tracking of objects in the macaque dorsomedial frontal cortex. bioRxiv 2022.06.24.497529 [Preprint] (2022). https://doi.org/10.1101/2022.06.24.497529.
40. R. Rajalingham, A. Piccato, M. Jazayeri, The role of mental simulation in primate physical inference abilities. bioRxiv 2021.01.14.42674 [Preprint] (2021). https://doi.org/10.1101/2021.01.14.426741.
41. A. Nayebi, R. Rajalingham, M. Jazayeri, G. R. Yang, "Neural foundations of mental simulation: Future prediction of latent representations on dynamic scenes," in *Proceedings of the 37th International Conference on Neural Information Processing Systems (NIPS '23)* (Curran Associates Inc., 2023), pp. 70548–70561.
42. M. A. Goodale, D. A. Westwood, An evolving view of duplex vision: Separate but interacting cortical pathways for perception and action. *Curr. Opin. Neurobiol.* **14**, 203–211 (2004).
43. J. P. Gallivan, D. A. McLean, K. F. Valyear, J. C. Culham, Decoding the neural mechanisms of human tool use. *eLife* **2013**, e00425 (2013).
44. K. F. Valyear, C. Cavina-Pratesi, A. J. Stiglick, J. C. Culham, Does tool-related fMRI activity within the intraparietal sulcus reflect the plan to grasp? *Neuroimage* **36**, T94–T108 (2007).
45. R. E. B. Mruczek, I. S. von Loga, S. Kastner, The representation of tool and non-tool object information in the human intraparietal sulcus. *J. Neurophysiol.* **109**, 2883–2896 (2013).
46. J. Duncan, The multiple-demand (MD) system of the primate brain: Mental programs for intelligent behaviour. *Trends Cogn. Sci.* **14**, 172–179 (2010).
47. E. Fedorenko, J. Duncan, N. Kanwisher, Broad domain generality in focal regions of frontal and parietal cortex. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 16616–16621 (2013).
48. J. Fischer, B. Z. Mahon, What tool representation, intuitive physics, and action have in common : The brain's first-person physics engine. *Cogn. Neuropsychol.* **38**, 455–467 (2022).
49. A. Mitko, A. Navarro-Cebrián, S. Cormiea, J. Fischer, A dedicated mental resource for intuitive physics. *iScience* **27**, 108607 (2024).
50. J. B. Julian, E. Fedorenko, J. Webster, N. Kanwisher, An algorithmic method for functionally defining regions of interest in the ventral visual pathway. *Neuroimage* **60**, 2357–2364 (2012).
51. J. V. Haxby, M. I. Gobbini, M. L. Furey, A. Ishai, J. L. Schouten, P. Pietrini, Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science* **293**, 2425–2430 (2001).
52. A. M. Dale, Optimal experimental design for event-related fMRI. *Hum. Brain Mapp.* **8**, 109–114 (1999).
53. D. M. Bear, E. Wang, D. Mrowca, F. J. Binder, H.-Y. F. Tung, R. T. Pramod, C. Holdaway, S. Tao, K. Smith, F.-Y. Sun, L. Fei-Fei, N. Kanwisher, J. B. Tenenbaum, D. L. K. Yamins, J. E. Fan, Physion: Evaluating physical prediction from vision in humans and machines. arXiv:2106.08261 [cs.AI] (2021).
54. T. A. Carlson, J. Brendan Ritchie, N. Kriegeskorte, S. Durvasula, J. Ma, Reaction time for object categorization is predicted by representational distance. *J. Cogn. Neurosci.* **26**, 132–142 (2014).
55. S. Thorat, D. Proklova, M. V. Peelen, The nature of the animacy organization in human ventral temporal cortex. *eLife* **8**, e47142 (2019).

Acknowledgments: We thank K. Brewer for helping with stimulus creation for experiment 1, G. Woo for helping with deep learning model analyses, and members of the Kanwisher laboratory at MIT for valuable comments and feedback on the manuscript. **Funding:** This work was supported by NSF grant 2124136 to N.K. and NSF Science and Technology Center—Center for Brains, Minds, and Machines grant CCF-1231216. **Author contributions:** Conceptualization: R.T.P., E.M., N.K., and J.B.T. Investigation: R.T.P., E.M., and C.X.F. Methodology: R.T.P., E.M., and N.K. Data curation, formal analysis, and visualization: R.T.P. and E.M. Writing—original draft, validation, and project administration: R.T.P. and N.K. Writing—review and editing: R.T.P., E.M., N.K., and J.B.T. Supervision and funding acquisition: N.K and J.B.T. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. Stimuli and data used in the paper can be downloaded from 10.5061/dryad.2547d7x35.

Submitted 15 July 2024
Accepted 25 April 2025
Published 30 May 2025
10.1126/sciadv.adr7429