Representations in the Intuitive Physics Engine (IPE)

Shashank Srikant, shash@mit.edu

Summary.
Key questions asked. Are different physics concepts distinctly represented in the IPE regions of the brain? Can a decoder be designed using these representations to mimic the output of a computational physics engine?

Experiment 1 (preliminary investigation). Are different physics concepts like stability, motion of objects, etc. represented uniquely in the IPE?
Methods used. RSA

Experiment 2 (follow up of Exp 1). Can IPE representations of motion and stability of objects in 'what next' scenes be decoded to predict what actually happens next?
Methods used. Voxel information decoder, deep learning

Introduction.

Fischer et al. [1] identified regions of the brain termed the *Intuitive Physics Engine* (IPE). These regions are sensitive to physical information present in a scene, and are likely involved in scene understanding and performing physics inference. In other work, through different psychophysical tasks, Battaglia et al. [2] further showed that an IPE can be modeled as computational physics engines used in computer graphics programs.

A generic physics engine, such as the one demonstrated in [2],[1] typically comprises a simulator for rigid body dynamics, which plays out the effect of a wide variety of forces like gravity, friction, viscosity, etc. on an object in different weight and motion settings. In practice though, applications utilizing physics engines tend to be narrowly defined. A physics engine for a golf game would be parameterized differently than a physics engine for a game of pong or a game of computerized Jenga.[2] A 'golf engine' will have air drag and cross-wind as important input parameters besides the initial conditions of the ball, while a "pong engine" would focus on the physics of elastic collisions more than forces of friction and drag; a "Jenga engine" would focus on the centers of masses of the stacked objects.[3] This modular approach to defining variants of physics engines, each focusing on a narrow set of rules of Newtonian physics, suggests that they are largely non-overlapping, functionally distinct ways of characterizing the physical world around us.

If it is indeed convenient and efficient to process such "task types" differently in computational physics engines, does the IPE in our brains do so as well? Are there different sub-regions within the identified IPE which are responsible for different tasks like inferring motion and inferring centers of mass? Do the IPE regions represent each of these task types differently, implying the existence of a neural mechanism which handles such information differently?

Experiment 1. We first investigate whether the regions identified as IPE in [1] represent tasks governed by different aspects of physics uniquely. The first experiment uses RSA

[1] https://chandlerprall.github.io/Physijs/examples/jenga.html
[2] Open Dynamics Engine, http://ode.org/
[3] See informal discussion here - https://gamedev.stackexchange.com/questions/129686/what-exactly-is-a-physics-engine

(Representation Similarity Analysis) to determine if the activations detected by fMRI in the IPE on stimuli presenting different physics tasks are similar to hand-labeled tags provided to the stimuli.

We investigate two physical properties in this experiment – stability and motion.

- ***Stability*** refers to whether the net forces on an object is zero, which consequently determines if the object shall remain stationary. The outcome of interest in this condition is whether an object would remain at rest and not end up crashing or tumbling. We note here that the underlying physics which results in an object remaining at rest may vary depending on the particular setting of the environment. We investigate two such settings as separate sub-conditions. Details below.
- ***Motion*** here refers to an object's kinematics in a 2-dimensional plane.

Other physical properties like fluid dynamics, collisions, rotational acceleration, etc. can be investigated in future work.

Hypotheses. The following are the possibilities -
H1. The IPE has no distinct regions which are sensitive to either the physics of motion or the physics of stability.
H2. The IPE has a region sensitive only to motion but not stability.
H3. The IPE has no region sensitive to motion, but has a region sensitive only to stability.
H4. The IPE has regions sensitive to both, motion and stability.
H5. The IPE has distinct regions sensitive to both, motion and stability.

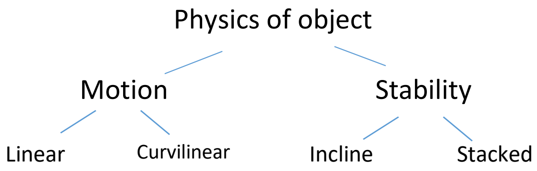
Stimuli design and details.

- To keep the entire analysis simple, this experiment focuses on scenes with only a single object being present in a scene. This analysis can be extended to multiple objects, but would then have to control for the effect of collision, which is a distinct physics phenomenon in itself.
- The physics of motion is described in two settings in the stimuli – linear, and curvilinear.
    - ***Linear motion*** refers to a body's motion whose locus is a straight line.
      e.g. the motion of a runner running on a straight road.
    - ***Curvilinear motion*** refers to a body's motion whose locus is a part of a circle.
      e.g. the trajectory of a ball thrown from a cliff.
- The physics of stability is described in two settings – stacked, inclined.
    - ***Stacked stability*** refers to the condition when the net forces on the center of mass of a system of objects is zero. Although this physics of this is best illustrated with multiple bodies stacked one atop other, we choose the situation where an object is placed at the tip of surface like a table.
    - ***Stability on an incline*** refers to the condition when the net forces acting on an object placed on an inclined plane is zero. Key concepts in physics determining such a stability are
        - Incline of the plane
          e.g. a higher incline increases the component of gravity acting on the body, and hence will cause it to roll or tumble.
        - Mass, and consequently the moment of inertia of the object
          e.g. a spherical object will tend to roll down a plane no matter how small the incline. A box will not.
        - Friction of the surface.
          e.g. an icy surface will cause even a box on a small incline to slide down.

      In this work, we focus only on varying the incline to depict stability or the lack of it. The surface in each scene is a wooden wedge, and the object is a cube.

See subsection ‘Content design’ and ‘Discussion’ for a note on the differences between these two sub conditions that might affect this experiment.

- This experiment can be carried out with 2x2 unique items (with repetitions), and in principle be extended to any number of such categories. The current design can be illustrated as -

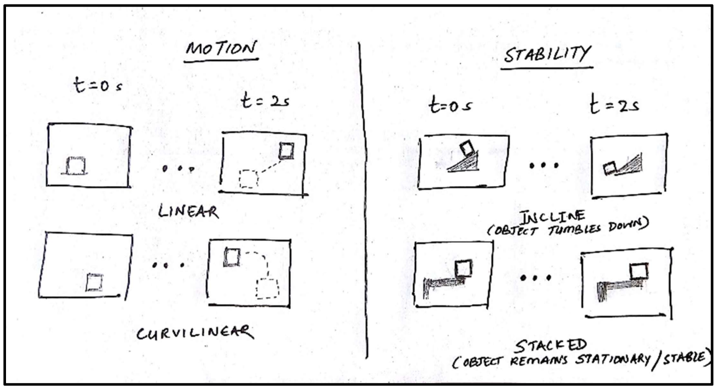


- The stimuli can be presented in two formats – statics images, or very short video clips. Short video clips have these advantages –
  - They can depict the physics of motion naturally. Static images would need traces or hashed lines to depict motion.
  - Videos of stability will involve objects falling or tumbling, which also involves motion. This will allow us to then compare the two conditions – Motion vs. motion induced by instability, and check for a possible confound.

  We hence choose to use 3 second visual clips.
- **Content generation.** There exist three choices for content generation –
  - Manually design animated versions of the content using a graphics editor.
  - Source videos from YouTube or other popular sites and edit and ensure they fit the requirement.
  - Shoot videos in a simple setup using a decent camera.

  Option 1 will likely be the most convenient. It will also allow for easy control over variability and noise in the depicted scenes.
- **Content design.** All the main content will be placed at the center of the screen, to minimize eyeball movement. To control for the shape of the object, a cuboid/box is used in both, *Motion* and *Stability* stimuli.

  In *Motion* stimuli, the three seconds interval will depict a box moving from one point to another, either along a line or a curve. The direction of movement is randomized. This is done to see if the concepts generalize across such minor variations. A box moving from the top to bottom of the screen, or from left to right still signifies motion. We investigate whether this abstraction is captured in the representations.



  In *Stability* stimuli, the two second interval will start with an object either at the edge of a surface (stacked), or at the top of a wedge (incline). In half the cases, the object falls down or tumbles down, and in the other half remains stationary.

  Important. As demonstrated in the subsequent experiment, we design these stimuli such that there is distinct activity that happens in the last 1 second of the video. For the motion stimuli, this pertains to the box simply following through with the trajectory shown in the first two seconds. For the stability stimuli, this crucially pertains to the object falling or tumbling during this last one second (in the case it is presented in an unstable scenario).

  - The angles of the incline are sampled from the set *theta* = {10, 30, 70} degrees.
  - The distance of the edge of the box from the tip of the table top is sampled from the set *len* = {L-0.1, L/2, L/5}, where L is the edge length of the box. Lesser the distance, higher its stability. The stacked setting in the image below corresponds to L/5.
  - When *len* = L-0.1 (right at the edge of the table), and when *theta*=70 (steep wedge incline), the condition of stability is conveyed by showing the box to jitter and seemingly just managing to balance itself.
- **Number of stimuli.** The total number generated for the experiment are
  - # repetitions for each condition: 20 (will get a second opinion from advisor on whether this number will be sufficient).
  - # unique (Motion, Linear) items: 6 (with differing start positions and direction of uniform motion)
  - # unique (Motion, Curvilinear) items: 6 (with differing start positions, and arc lengths of the uniform motion)
  - # unique (Stability, Inclined) items: 3 + 3 (for each angle of the wedge, equal number of stable and unstable situations).
  - # unique (Stability, Stacked) items: 3 + 3 (for each length from tip, equal number of stable and unstable situations).
  - Total items = 20 x (6 + 6 + 6 + 6) = 480
  - Each video takes 3 seconds; Hence, experiment scan time = 480 x 3 = 24 minutes. (This is a lower bound. See below for total scan time.)
  - To keep it simple, repetitions will be exact replicas of the 24 unique items.
- **Stimuli presentation.**
  - Each run will contain 24x2 = 48 items, i.e. one set of unique items + one set of its repetitions.
  - Hence, total number of runs = 480 / 48 = 10.
  - This is modeled as a one-back test to ensure the participant is paying attention. At the end of each stimulus, show a blank screen for 3 seconds, during which time the participant presses either A or B to signify whether the stimulus she just saw was the same as the one preceding it or not, respectively.
    - Note – Can perhaps reduce the 3s time window to something lesser. Unsure.
  - The order of the 48 items in a given run is quasi random, to ensure there are at least a few items which pass the one back test.
  - Time per run = (3+3) seconds x 48 items = ~5 minutes.
  - Total time = (Time per run x Number of runs) = 5 x 10 = ~50 minutes.
  - One break midway can be provided to the participant.
  - The total experiment time will be 50 minutes + the amount of time takes to run the stimuli in Fischer et al.

Experiment procedure.

- Contact the authors of Fischer et al. [1] and obtain their stimuli.
- Run the stimuli used in experiment 1 and/or 2 to localize the ROI identified in their work.
- Run the current design of 480 items spread over 10 runs.

- We also run the eye saccade experiment as in [1] to account for eye movement in the ROI.

Data analysis and preparing base similarity matrices for RSA to compare against.

- Obtain the activations in the ROI identified in [1] and replicate results of experiments 1 and/or 2.
- All analysis of the current design will first be done in these identified regions.
- Divide the identified regions of interest into sub-regions of an appropriate dimension.
- For all voxels in a sub-region whose activity was measured during the time period when the participant was shown the stimulus (and not when she responded to the one-back test), obtain the correlation between every two conditions in this experiment. Average it out by the number of replications. Calculate the average correlation (absolute values) for the entire set of voxels and represent it as a 48x48 matrix as show in Figure 2. **We refer to this matrix as F_k**, where the subscript denotes sub-region *k*.
- This provides the representation matrix along the 2 main conditions (Motion, Stability) and 2 sub conditions each (Linear, Curvilinear; Stacked, Inclined).
- Such a matrix would need to be compared against a reference matrix. We obtain these reference matrices by showing the stimuli to subject matter experts (SME). In our case, these are two physics grad students at MIT. For each stimulus, we ask the SMEs to answer five questions (details below) and get them to answer them on a scale of 1-10. After this exercise, the inter-rater correlation is verified, and if sufficiently high (>0.85), we proceed with our analysis. The five reference matrices are then simply the correlation of their average response across the four conditions.

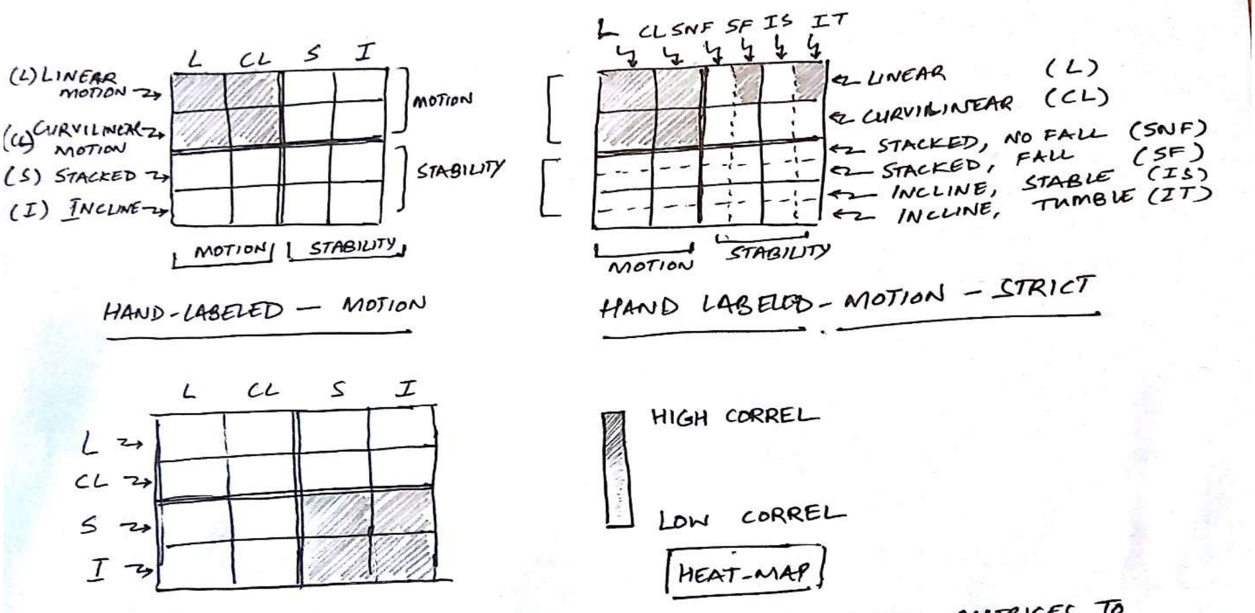


FIG. EXPERT-LABELED CORREL MATRICES TO COMPARE FMRI MATRICES AGAINST

- We ask the following questions to the SMEs –
    - **Baseline 1 (B1).** The stimulus depicts the concept of *Motion.* Rate 1-10.
    - **Baseline 2 (B2).** The stimulus depicts the concept of *Stability.* Rate 1-10.

    - **Hand labeled – Motion (HLM).** The entire length of the video contains an object in uniform motion in a 2D plane, with net resultant force being zero. Rate 1-10.
    - **Hand labeled – Motion – Strict (HLMS).** The video contains, at some point, an object in motion in a 2D plane. Rate 1-10.
    - **Hand labeled – Forces (HLF).** The video contains an object whose stable state position is affected by the resultant force acting on its center of mass. Rate 1-10.
- We expect HLM, HLMS, and HLF to be represented as shown below. Further, we expect B1 to correspond to HLM and B2 to correspond to HLF respectively. In case it does not, we could proceed with HLM, HLMS, and HLF for the remainder of our analysis.
- We create a 48x48 matrix **Random** containing random correlations sampled uniformly between [0.1, 0.8]. This serves as a baseline.
- In the illustration of expected hand labeled matrices, the reason for HLMS having a higher correlation in the top two boxes in the right quadrant is that the process of falling and tumbling may be perceived as motion. If there exist regions sensitive only to motion, then the fMRI voxel correlation matrix will be similar to HLMS over HLM.
- We started with five possibilities regarding sub regions in the ROI (refer to subsection *Hypotheses* at the beginning of this section). We enumerate the possibilities of data we expect to see against each such hypothesis.
    - If H1 is true, then the fMRI matrix for a sub-region *k*, F_k, will not correlate with any of HLM, HLMS, HLF, or HLB. It will rather correlate with Random.
    - If H2 is true, F_k will correlate with HLMS most likely, over HLM. It would be an interesting case if it correlates with HLM > HLMS. Also, it will not correlate with HLF.
    - If H3 is true, F_k will correlate highly with HLF but not HLMS or HLM.
    - If H4 is true, F_k will correlate highly with HLMS/HLM and HLF.
    - If H5 is true, F_k will correlate highly with HMLS/HML and F_s will correlate with HLF, for two different sub-regions *k* and *s* in the ROI.
    - Another possibility is that all sub-regions regions *k* more or less are highly correlated with both, HLMS and HLF.
- **Other analyses.** Other aspects such as the sensitivity of regions to the angle of the wedge, how far the box is placed from the edge of the table, direction of motion, etc. can be measured as well through a similar RSA analysis. This will require getting SMEs to create hand-labeled representation matrices as before, quantifying these measures of interest.

Discussion.

Alternate explanations.

- If the data turns out in a way that there exist sub-regions sensitive to only a fraction of the stability items, a possible explanation could be the difference in the physics being applied in the two stability sub conditions. While the incline condition relies on friction and angle of incline, the stacked relies purely on center of mass imbalance. This could further be investigated by getting the SMEs to reevaluating the HLF matrix with this specific distinction in mind to see if that really explains the fMRI data.

- If the RSA analysis points to H5 being true (i.e. there exist distinct sub-regions sensitive to motion and stability respectively), the simplest possible explanation could be that the stability stimuli had more objects than the motion stimuli. Notice, in each of our stability stimuli, we place the object on a wedge or a table. This region could simply be sensitive to the number of objects and not the physics of stability of an object. We did consider stimuli which cleverly demonstrated object stability without it having to rest on any surface. For example, this top can be shown to topple without a reference. However, we wanted to cover a wider variety of conditions that we see every day on common objects.

  If this is the only alternate explanation feasible, a simple follow up experiment can be conducted to identify the sensitivity of this specific sub-region to differing number of objects in a scene.

If on running additional experiments to account for the alternate explanations mentioned above, we find that H5 continues to be validated by our data, we are encouraged to consider the possibility that regions indeed are specialized to contain information regarding motion and stability respectively.
We design the following experiment as a stronger test to verify if these regions indeed store specific information on motion and stability.

Experiment 2.
Motivation.
If we find separate regions in the IPE which are sensitive to the physics of motion and the physics of stability respectively, a natural question that follows is what information is really being represented in these regions.

To contrast the working of this region, consider a computational physics engine designed to model the motion of an object. It would consider as inputs the current velocity (position, speed) of the object and the forces acting on it. For a given time instance T, it will internally compute how the laws of motion would affect its velocity. The actual algorithm could be probabilistic (as suggested in [2]) or determinstic. It then outputs its prediction. If we were to model the IPE's sub-regions for motion similarly, we then expect the region to be the component which computes the final position of the modified object. It should then be possible to decode this information to perform motion-related predictions. We can make a similar argument for predicting stability information.

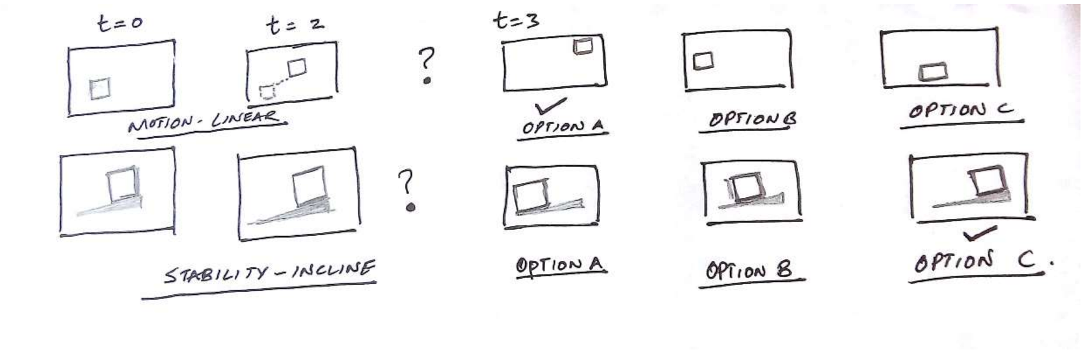
Hypotheses.

- **H1.** Information in the motion regions cannot be decoded to predict motion. Similarly, information in the stability regions cannot be decoded to predict stability conditions.
- **H2.** Information in the motion regions can be decoded to predict motion. Information in the stability regions are able to predict stability conditions. Moreover, information in the stability regions cannot be decoded to predict motion. Similarly, information in the motion regions cannot be decoded to predict stability.

- **H3.** Other two permutations between what information can be decoded and predicted.

Experiment design.
This experiment can be designed in two ways.

- The first is to simply tweak data from the previous experiment to fit the requirements of this experiment. In this experiment, want to consider the representation which likely pertains to the subject having internally computed and utilized her IPE to predict what happens next in the scene. One approximation could be to consider the neural activity in the first 2 seconds of the video only, since by the stimuli design (see subsection *Content Design* in the previous experiment), the critical activity, if any, happens only in the last one second. Hence, if our hypothesis of what is represented in this region is correct, the representation formed at the end of the first two seconds ought to be sufficient to predict what happens at the end of the third second.
- The other way is to present the same stimuli as the previous experiment, but stop the video at the end of 2 seconds. Instead of presenting this as a one-back task, present four options for the possible final state, and ask the participant to endorse one of them. An illustration of the stimuli and expected response –

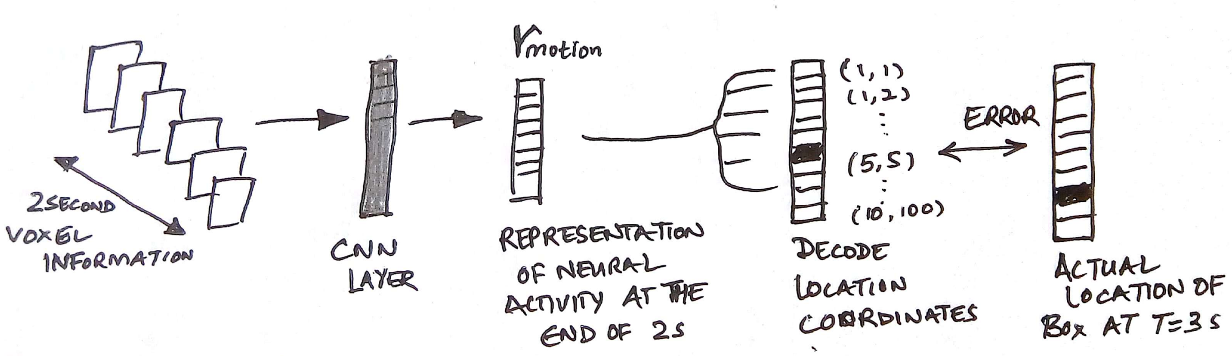


- **Discussion on design choice.**
The second design presents a cleaner structure tailored to elicit specific predictive information at the end of the third second. However, it also suffers from the risk of attention being paid to the specific task, as a consequence of which the neural activity in these regions may contain this predictive information. Even if this were the case, it would be an interesting result.

  For the purposes of this experiment and write up, we proceed with the former choice of analyzing information present at the end of the 2nd second. We will subsequently run the other design choice as well.

Experiment details.

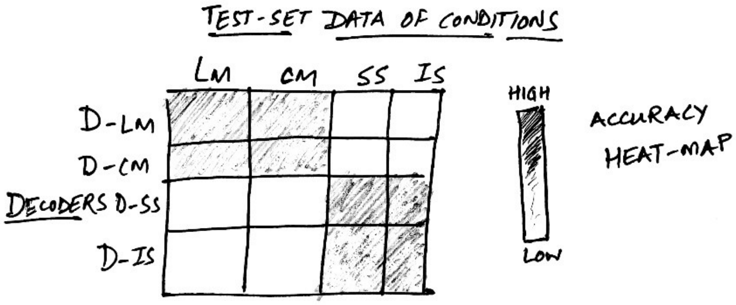
- Since we proceed with a post-hoc analysis of the information captured in the previous experiment, we do not need to test any subjects on any stimuli.
- If we were to implement the latter design choice, every configuration decided in the previous would remain the same barring what the subject does – in this scenario, she will be required to guess the most likely configuration at the end of the third second from three choices shown to her after the stimulus ends. This does not affect our analysis since we look at activation data pertaining to only the two seconds when the stimulus was shown to the subject, and not when the subject was responding to the behavioral task.


2SECOND VOXEL INFORMATION
CNN LAYER
r_{motion}
REPRESENTATION OF NEURAL ACTIVITY AT THE END OF 2s
(1,1)
(1,2)
(5,5)
(10,100)
DECODE LOCATION COORDINATES
ERROR
ACTUAL LOCATION OF BOX AT T=3s

MOTION REGION DECODING

Analysis.

- We split the data into an 80-20 train and test set across each of the four sub-conditions.
- We train a decoder for each of these sub conditions separately. They are annotated as – D-LM (linear-motion), D-CM (curvilinear-motion), D-SS (stacked-stability), D-IS (inclined-stability)
- We design decoders using the architecture shown above. Each decoder will be input the voxel frames until t=2 seconds. A CNN layer would then be trained to produce a representation vector of the activity capturing all the information in the voxel frames. The signal to train the CNN would be the error incurred in decoding the predicted position of the object. For the stacked-stability stimuli, the decoder can be replaced with a binary classifier of whether the object would fall or not. For the rest, the positions from both, the wedge (stability) and the 2-D plane (motion) can be linearized to be represented as a 1-D vector of coordinates.
- Once the classifiers are trained, any data point pushed into the decoder will produce an intermediate representations which capture the average neural activity at the end of 2 seconds – R-CM, R-LM, R-IS, R-IS.
- On the test set, we pass the each conditions' dataset into each decoders and measure accuracies. We expect a 4x4 table of accuracies.



Discussion.

- If the hypothesis is true that the regions indeed serve as a physics engine, and compute physical properties of objects presented in a scene, we would expect the decoder accuracies to be tabulated as above. Here, we observe that the decoders LM and CM generalize enough to accurately classify motion regardless of it being linear or curvilinear. Likewise, the stability decoders predict irrespective of which stability related condition is provided as input. Such a result also demonstrates exclusivity, in that information processed in the stability regions do not contain any information about motion and vice versa.

- Such accuracy results would also settle the alternate explanation raised in the previous experiment. This data would clearly show that there is meaningful and exclusive information being stored in the stability regions of the IPE. To further test it, another decoder can be trained to predict the number of objects in a scene, and SS, IS data can be fed to it to see if it predicts such information well.
- In case this hypothesis was not fully true, and instead linear information was computed differently than curvilinear information, then the CM--D-LM square and LM—D-CM square would show low accuracy information. The same would follow for the stability regions as well.

Conclusion.

This work proposes two experiments to answer whether different physics operations are processed differently in the IPE, and if they are, what information such regions contain. This work leads to other important questions regarding the IPE such as

- whether there exist other regions for other physics operations,
- how do these regions communicate with each other,
- whether affecting these regions using TMS would impede internal simulation
- do these regions have a spatial map, akin to a retinotopic map, which is sensitive to the range of inputs for the operation it specializes in?

References.

[1] Fischer, J., Mikhael, J. G., Tenenbaum, J. B., & Kanwisher, N. (2016). Functional neuroanatomy of intuitive physical inference. Proceedings of the national academy of sciences, 113(34), E5072-E5081.

[2] Battaglia, P. W., Hamrick, J. B., & Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. Proceedings of the National Academy of Sciences, 110(45), 18327-18332.

Note - This was an initial idea which I started working on, which later was shelved. The goal was to stick to simple behavioral experiments to discover some interesting information about the IPE. Only later did I find out there's a related work which addresses parts of this question.

Specifying the architecture of an Intuitive Physics Engine (IPE)

Shashank Srikant, shash@mit.edu

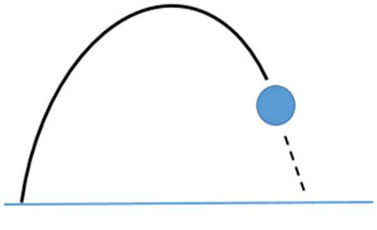
Introduction.

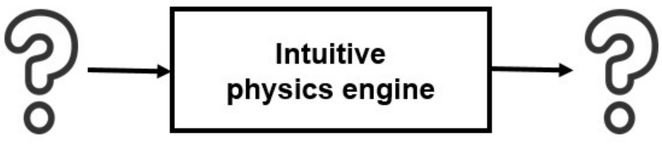
Humans are hypothesized to be equipped with an *intuitive physics engine* (IPE). Through different psychophysical tasks, Battaglia et al. [1] showed that an IPE can be modeled as physics engines used in computer graphics programs. Quoting from their work, "these models use approximate, probabilistic simulations to make robust and fast inferences in complex natural scenes where crucial information is unobserved".

While this work attempts to model the algorithmic aspects of an IPE, the question of what the overall architecture of this engine might be, and what it might be parameterized by are left unexplored. A commercial physics engine, such as the one used in [1], typically allows to overspecify the set of properties associated with each object in a scene, such as its coordinates, angular momentum, instantaneous acceleration, etc. Likewise, it computes values for each such property at the end of its simulation run. Is our IPE as resourceful? Can we access values of various properties associated with an object at the end of our "mental simulation" of a scene, or is it restricted to a specific set?

Fully specifying the engine, i.e. knowing what its inputs and outputs possibly can be, will likely reveal the nature of physical tasks we can intuitively reason about. Are we adept at predicting just about any property of an object in a physical scene, or are there limited properties we are capable of inferring. For instance, when shown a ball moving in a curvilinear motion, we seem to quickly be able to infer the location of the ball at any point in the future. But shown the same image, can we predict other properties of the ball, such as its instantaneous speed, or its acceleration? Similarly, it is not clear what inputs the IPE is parameterized by. Are common forces of nature acting on a body, such as gravity and friction, already "baked into" the IPE, or are they explicitly input to the system? Can the IPE accommodate and learn new values for these forces when subject to different environments?

Consequently, answering this will allow us to understand which tasks can even be labeled as those requiring "intuitive physics". Current work does not attempt to answer this question either.




?
Intuitive physics engine
?