Humans and convolutional neural networks prioritize similar visual features in intuitive physics judgments

Ren Calabro (rencalabro@uchicago.edu)
Department of Psychology, University of Chicago
5848 S. University Avenue, Chicago, IL 60637 USA

Kannon Bhattacharyya (kannon@uchicago.edu)
Department of Psychology, University of Chicago
5848 S. University Avenue, Chicago, IL 60637 USA

Wilma Bainbridge (wilma@uchicago.edu)
Department of Psychology, University of Chicago
5848 S. University Avenue, Chicago, IL 60637 USA

Yuan Chang Leong (ycleong@uchicago.edu)
Department of Psychology, University of Chicago
5848 S. University Avenue, Chicago, IL 60637 USA

Abstract

Humans reliably infer complex physical relationships between objects in everyday scenes, yet the mechanisms underlying these judgments remain unclear. We explored whether convolutional neural networks (CNNs) can approximate intuitive physical reasoning by capturing statistical regularities in visual experience. We trained a CNN (Inception-v4) to predict tower stability and tested how well its outputs aligned with human judgments (N = 500). CNN predictions more closely matched human judgments (r = 0.718, p < 0.001, accuracy = 81%) than ground-truth predictions from physics simulations (r = 0.406, p = 0.002, accuracy = 68%), suggesting that both CNNs and humans rely on visual heuristics. Eye-tracking data revealed that CNN importance maps overlapped significantly with human gaze patterns, indicating shared attention to features statistically predictive of physical outcomes in intuitive physical judgments. Our findings show that CNNs trained on visual data capture perceptual cues used in human intuitive physics, highlighting their value as models of heuristic reasoning.

Keywords: eye tracking; statistical learning; artificial intelligence; vision; neural networks

Introduction

We form rapid and robust intuitions about our physical environments during everyday decision-making, such as the stability of a stack of dishes, the trajectory of colliding billiard balls, or the stopping time of a car in traffic. However, complex, ambiguous, or novel scenarios may result in perceptual errors (Kelley & Kelley, 2014). Thus, understanding how we reason about our surroundings is crucial for identifying the limitations of our intuitions. In turn, this may help people make more accurate judgments about their environments and become better decision-makers.

Despite its foundational role in human intelligence, the cognitive processes underlying intuitive physical inference are not fully understood (Ludwin-Peery et al., 2021). One prominent theory suggests that humans make judgments about physical scenarios by mentally simulating them, using internal models of physics to support flexible and dynamic inference (Battaglia, Hamrick, & Tenenbaum, 2013). This process is often likened to how video game engines simulate real-world physics to predict outcomes. However, in fast-paced or uncertain situations, individuals may instead rely on perceptual heuristics– learned shortcuts that support efficient decision-making without requiring full simulation (Callaway, Hamrick, & Griffiths, 2017). While generally effective, heuristics can introduce systematic biases, such as an overreliance on visual cues like symmetry or balance when judging an object's stability (Tversky & Kahneman, 1974). These heuristics may arise through statistical learning, whereby individuals extract regularities from past experience to make rapid judgments in novel contexts. Such learning supports the use of cognitive shortcuts that complement simulation by enabling quick and relatively accurate inferences. Recent work has shown that human judgments of object stability can be predicted from visual features, further supporting the role of perceptual heuristics in intuitive physical reasoning (Sanborn, Mansinghka, & Griffiths, 2013; Liu, Ayzenberg, & Lourenco, 2024).

In recent years, advances in artificial intelligence have provided a powerful tool for investigating human physical reasoning (Wu et al., 2015). Convolutional neural networks (CNNs), trained on large-scale image datasets, have demonstrated impressive performance in diverse tasks, including object recognition, saliency prediction, and physical inference (Dosovitskiy et al., 2016; Lerer, Gross, & Fergus, 2016; Groth et al., 2018). Because CNNs learn statistical patterns directly from visual input and do not have explicit access to physical laws, they offer a compelling framework for examining the kinds of perceptual heuristics that may underlie intuitive physics. This contrasts with simulation-based cognitive models such as the Intuitive Physics Engine (IPE), which posit that humans internally simulate physical events using probabilistic models (Battaglia et al., 2013; Zhang et al., 2016). Although the IPE provides a cognitively plausible account of mental simulation, our study instead uses a physics engine as a benchmark to establish a ground-truth reference for physical outcomes. Rather than directly modeling physical dynamics, we evaluate how closely CNNs approximate human judgments and where both diverge from physical truth. In doing so, we explore the extent to which perceptual heuristics align with or deviate from actual physical stability.

Notably, prior research has shown that CNN-based models can predict human visual attention across a variety of contexts (Kümmerer et al., 2015; Borji & Itti, 2013), and that

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

perceptual features alone can, in some cases, approximate human stability judgments (Conwell, Doshi, & Alvarez, 2019). Unlike traditional simulations that rely on predefined physical rules, CNNs simulate statistical learning processes directly from raw data, capturing complex regularities that emerge from visual experience. This makes CNNs particularly well-suited for modeling how humans form judgments about stability and other physical properties through visual experience.

In this study, we investigate whether CNNs can predict human intuitive physical judgments about stability. We hypothesize that CNNs trained on stability prediction tasks will produce predictions consistent with those of human participants. To explore whether humans and CNNs rely on similar visual features when making stability judgments, we used eye-tracking and occlusion studies to map attention patterns and identify the visual features most critical for assessing stability. In addition to examining shared visual features between humans and CNNs, the eye-tracking analysis assessed whether divergent human stability judgments arise from differences in how visual information is sampled (Hayhoe & Ballard, 2005). To test these hypotheses, we first (1) trained a CNN to predict tower stability and assessed how well its predictions aligned with human judgments. Next, we (2) collected eye-tracking data as participants judged the stability of block towers and (3) compared human fixation patterns with model-generated importance maps. Our findings reveal that both human and CNN physical reasoning prioritize similar information, suggesting that intuitive physical judgments may rely on the identification of statistical regularities in the visual environment. Specifically, the results suggest that both CNNs and humans attend to features that are statistically predictive of stability, indicating that human intuitive physical judgments may be influenced by visual cues similar to those the model identifies through its training.

Methods

Stimuli

Our stimulus set was derived from the ShapeStacks dataset, which contains 20,000 block tower scenes ranging from 2 to 6 blocks tall, rendered from 16 camera angles (Figure 1a). While block towers of various heights were used for model training, we restricted the stimuli presented to human participants to 550 5-block towers to maintain consistency in visual scale and tower structure for subsequent eye-tracking and spatial analyses. We labeled each tower as "unstable" if at least one block fell under the influence of gravity in a physics engine simulation, and "stable" if not. Unstable towers were constructed by offsetting a single block's center of mass (2nd through 5th block), while the base block always remained stable.

Task design

Study 1: Prolific The block tower paradigm is classically employed in intuitive physics research and can be framed as a binary choice ("fall"/"stand") task (Figure 1b). In a pre-registered study, we recruited adult participants via Prolific for an online experiment (N=500). They were given a 5-cent bonus for correctly judging whether each block tower would fall or remain standing under the influence of gravity.

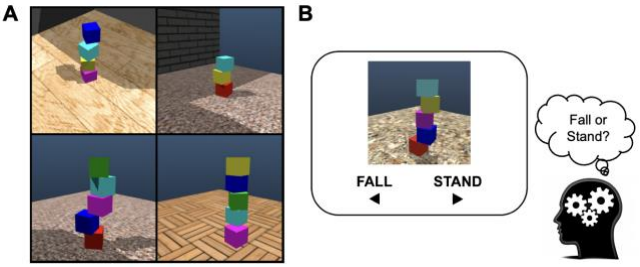

Figure 1: A. Images from the stimulus set presented to human participants and set aside for the CNN's test set. B. The task design.

The images were shown in random order. After receiving both verbal and written instructions and completing a practice round with feedback, participants were shown each image for 4 seconds and were prompted to respond whether the tower would "Fall" or "Stand" using the left or right arrow key. If no response was made within the 4-second timeframe, the trial would time out, and the participant would receive the message "Sorry, too slow! Please press the space bar to continue." After each recorded trial, the participant rated their confidence in their judgment on a scale from (1: "Very Unconfident", 2: "Somewhat Unconfident", 3: "Neither Confident nor Unconfident", 4: "Somewhat Confident", 5: "Very Confident"). Participants did not receive feedback about whether their response was correct, to avoid influencing future judgments and ensure that responses reflected their initial intuitions rather than learned strategies. Each participant saw 80 images out of 550 total; on average, each image was rated by 73 participants.

Study 2: Eye-tracking Participants (N = 40) were recruited from the University of Chicago and surrounding communities for an in-lab eye-tracking study. We used an EyeLink 1000 Desktop Mount to record participants' eye movements as they viewed a subset of 100 images from the Study 1 image set. The 100 images were selected to span the spectrum from clearly stable to ambiguous to clearly unstable based on average human ratings from a pilot study. The task structure was the same as for the Prolific study, with participants receiving a 5-cent bonus for each correct response.

Model training

CNN Architecture To investigate whether a convolutional neural network (CNN) could model human intuitive physical judgments about stability, we employed an architecture based on Inception-v4. This architecture was chosen for its ability to capture complex spatial and hierarchical features across multiple scales, making it particularly effective for tasks

involving detailed visual analysis. Importantly, nothing about the architecture itself is explicitly designed to encode principles of physics. Instead, the CNN gains its ability to predict stability judgments entirely through training on visual data, allowing it to learn statistical patterns from the input. Furthermore, Inception-v4 has previously demonstrated strong performance in predicting block stability in the ShapeStacks dataset (Groth et al., 2018).

To adapt Inception-v4 for the binary classification task (stable vs. unstable towers), we replaced the default softmax classifier with a custom head comprising a fully connected layer with 512 units (ReLU activation, L2 regularization) followed by batch normalization, dropout (rate = 0.2). The final output layer was a dense layer with a single unit and a sigmoid activation function, defined as:

$$\hat{y} = \sigma(z) = \frac{1}{1+e^{-z}} \quad (1)$$

where z is the output of the preceding dense layer. The sigmoid function maps z to a probability $\hat{y} \in [0,1]$, representing the likelihood that a tower is unstable. This design ensures that the model produces a probability suitable for binary classification.

Training and testing procedures The model was trained using the binary cross-entropy loss function, which penalizes incorrect predictions based on their confidence. The RMSprop optimizer was selected for its adaptive learning rate mechanism, promoting stable convergence during training. We employed a learning rate of $1x10^{-05}$ and trained the model over 80 epochs with a batch size of 32. Dropout was applied to mitigate overfitting, and model checkpoints were saved based on the lowest validation loss. The dataset was split approximately into 70% training, 15% validation, and 15% testing subsets (5,025 training towers, 1,076 validation towers, and 1,078 test towers). Towers varied in height from 2 to 6 blocks, and the 550 5-block towers used in the human experiments were entirely held out from model training, validation, and testing. Each tower was rendered from 16 distinct camera angles, and labels (stable = 1, unstable = 0) were derived from physics-based simulations from Groth et al. that provided ground-truth stability outcomes. To enhance generalization and mitigate overfitting, data augmentation techniques were employed during training. Specifically, random adjustments to the brightness, contrast, saturation, and hue of the images were applied. This approach helped the model learn more robust representations by exposing it to a wider variety of image variations. Model selection was based on validation loss, with training halted if the validation loss failed to improve for three consecutive epochs to mitigate overfitting.

Occlusion study

To assess the influence of different image regions on the CNN's stability predictions, we conducted an occlusion study to generate attention maps, revealing which areas of the scene the model prioritizes (Xu et al., 2015). This approach draws on methods from model interpretability that use input perturbation to assess the importance of different features based on changes in model predictions; that is, if occluding an area changes the model prediction, it indicates that the area is important for model prediction (Ribeiro, Singh, & Guestrin, 2016). For each image in the dataset, circular masks of varying radii (12, 16, 20, 24, and 28 pixels) were applied sequentially across the image, occluding specific areas. We then recorded the resulting changes in the model's prediction.

All images were first resized to 299 x 299 pixels, consistent with the input size used for model training. Background pixels were cropped out, and the block towers were centered and resized to a constant height of 200 pixels. This preprocessing step standardized the input images, ensuring that the results were not influenced by differences in the position or size of the tower within the image. For each radius, we created a sliding window approach with a dynamic stride size, set to half the radius to ensure overlapping coverage of adjacent regions, and a corresponding Gaussian blur applied with a sigma value equal to a quarter of the radius. The Gaussian blur was applied to soften the edges of the occluded region, simulating a more realistic transition rather than a hard boundary between occluded and unoccluded areas. The function is described as

$$G(x,y) = \frac{1}{2\pi\sigma^2} exp(-\frac{x^2+y^2}{2\sigma^2}) \quad (2)$$

where x and y represent the coordinates in the 2D space where the Gaussian kernel is being evaluated and σ is the standard deviation of the Gaussian kernel, which controls the spread of the filter.

At each window position, a circular mask was centered at the selected coordinates and applied to the image, and the resulting occluded image was fed into the CNN. The importance of the masked region was computed by subtracting the occluded prediction from the original prediction and taking the absolute change, with larger differences indicating greater relevance of that region to the model's decision. The final output of the occlusion study consisted of pixel-wise importance maps, indicating which regions of the image had the greatest influence on the CNN's stability judgments. These importance maps were used in subsequent analyses comparing model predictions to human fixation data.

Eye tracking data analysis

We aligned human fixation data with the resized, cropped, and centered images used in the occlusion study to ensure direct comparability between the human and CNN-based importance maps. Each subject's fixation data for a given image was represented as a 2D spatial distribution of fixations across the image. For each image, we concatenated fixation data across all subjects and applied a z-score normalization to the fixation counts at each pixel. This normalization ensured that the resulting human importance maps were on a consistent scale across images, reducing individual variability in fixation behavior.

To generate human importance maps, we followed the same methodology used for creating CNN-based importance maps. Specifically, the z-scored fixation values were treated as intensity values, and a Gaussian filter (Equation 2) was

applied to produce smooth, continuous maps of fixation density. We generated these maps using Gaussian kernels of varying sizes, corresponding to the five different radii applied in the occlusion study. The resulting human importance maps highlight regions where participants consistently fixated while making stability judgments. These maps allow for a direct comparison with the regions identified by the CNN as important for stability predictions, across each unique image and radius size.

By aligning human and CNN-generated importance maps in this way, we aimed to quantify the degree of overlap between human information-sampling strategies and the model's learned representations. This alignment allowed us to assess how closely the CNN's attention patterns correspond to those of humans when making judgments about physical stability.

Stability prediction comparison

Next, we compared the human and CNN-generated importance maps. We began by normalizing both the human and CNN-generated importance maps to ensure they were on the same scale for comparison. To assess the statistical significance of the observed pixel-wise correlation, we generated a null distribution through a permutation test. Specifically, we shuffled the image labels 10,000 times and calculated the mean correlation for each shuffle. This process created a null distribution representing the expected correlation between the human and CNN maps under the assumption of no true relationship, while taking into account the spatial correlation inherent in the data.

Results

A CNN predicts human stability judgments more accurately than physics engines

Before comparing model predictions to human responses, we first evaluated the reliability of human performance on the stability judgment task. Split-half reliability across raters was high (r = 0.978; Spearman-Brown corrected = 0.989), indicating strong internal consistency in stability judgments across participants. This suggests that the average human responses used for model comparison are a robust and stable measure of intuitive physical reasoning.

The trained CNN achieved up to 77.55% test accuracy in predicting whether towers were objectively stable or unstable, with performance across data splits averaging 74.91% (SD = 3.09%). This suggests that the model learned to make meaningful stability judgments aligned with ground truth physics, though generalization varied by split. Notably, this average slightly exceeds the 74.7% test accuracy reported by Groth et al. (2018) for a pretrained Inception-v4 model evaluated on a cube-only subset of the ShapeStacks dataset. In contrast, our model was trained entirely from scratch, demonstrating competitive performance without reliance on transfer learning. Our primary goal, however, was to test how well the CNN's predictions matched human judgments and whether the CNN attended to the same visual information that humans do when making stability predictions.

Remarkably, we found that the CNN's predictions were more strongly correlated with human judgments ($r = 0.718$, $p < 0.001$) than were predictions from the physics engine (point-biserial r = 0.406, p = 0.002). A dependent correlations test confirmed that this difference was statistically significant ($t(273) = 3.55$, p = 0.00023; Figure 2a). A non-parametric permutation test (100,000 iterations) further supported this result, showing that the observed correlation difference ($rCNN - rPE = 0.31$) was unlikely to have occurred by chance ($p < 0.001$). These findings suggest that the statistical patterns learned by the CNN from visual data align more closely with human reasoning than do explicit physical simulations. Importantly, the same physics engine was used to generate ground-truth stability values and compute point-biserial correlations with human responses, suggesting that the CNN may have learned to approximate the noisier human decision boundary rather than strict physical dynamics.

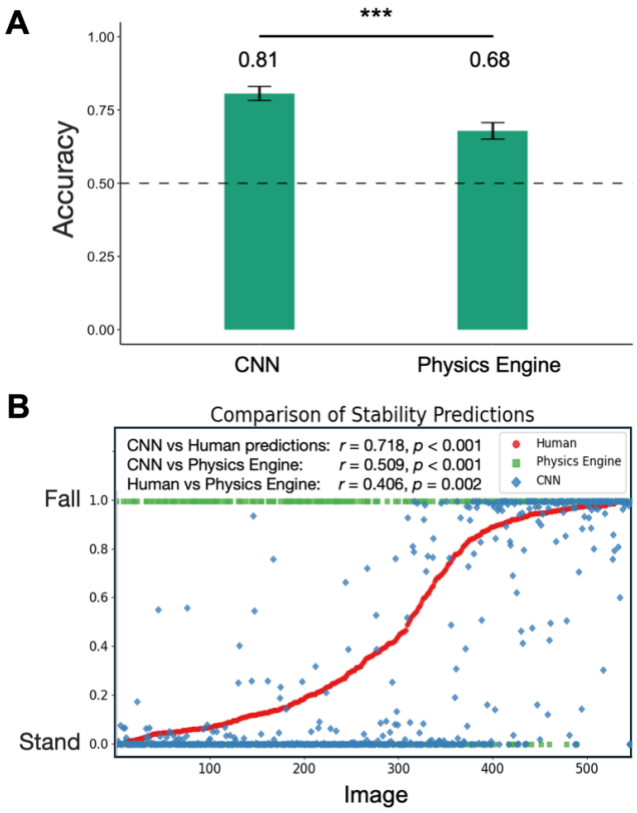

Figure 2: A. Accuracy comparison between the CNN and the physics engine in predicting human stability judgments. B. Image-wise stability predictions by humans (red), the physics engine (green), and the CNN (blue); physics engine predictions were binarized (stand = 0, fall = 1).

Interestingly, the CNN predictions had lower correlation with ground truth ratings than with human judgments ($r = 0.509$, $p < 0.001$; Figure 2b), indicating that the CNN, despite having been trained on the ground truth, better matched

human judgments. These results suggest that the CNN, trained on visual features alone, captures statistical regularities that align with human intuitions more effectively than physics-based models. This divergence suggests that the CNN may have internalized a decision boundary closer to that of humans than to the deterministic outputs of the simulation.

CNN and human visual attention patterns are significantly correlated

We also found significant correlations between human and CNN-generated importance maps across all tested radius sizes. For each radius, the human and CNN importance maps were significantly correlated, as shown in Table 1. These findings suggest that both the CNN and human participants attended to similar regions of the image when making stability judgments, providing evidence for the shared visual attention patterns between humans and the model (Figure 3).

Table 1: Mean correlation and p-values for different patch radii.

Patch radius	Mean correlation	*p*-value
12	0.510	0.013
16	0.519	0.018
20	0.525	0.014
24	0.524	0.01
28	0.520	0.00999

These results reinforce the validity of using CNNs as a proxy for human stability judgments in the context of visual attention. The significant correlations across various radius sizes suggest that the CNN effectively captures key aspects of human visual processing during stability evaluations.

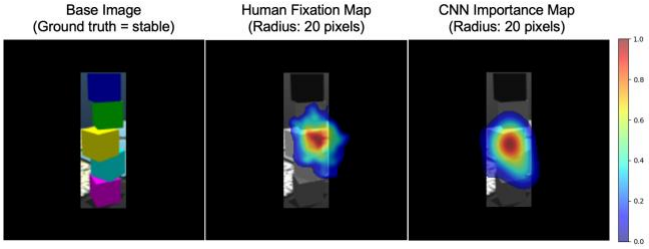

Figure 3: Comparison between human and CNN-generated importance maps for radius size 20. Heatmaps reflect model attention and gaze density, demonstrating overlapping visual priorities in stability estimation.

Eye-tracking reflects differences in information sampling

To assess whether eye-tracking reliably captured differences in information sampling in our task, we examined inter-subject correlation (ISC) in visual fixation patterns. For each image presented in the task, we computed fixation similarity between pairs of participants in two conditions: (1) *within-group* ISC, where both participants made the same stability judgment (e.g., both judged the tower as stable), and (2) *between-group* ISC, where participants made opposing judgments (e.g., one judged stable, the other unstable).

We then grouped comparisons by condition (within or between) and stability judgment (stand or fall). Mean ISC scores were computed for each participant in each condition, and paired t-tests were used to test whether within-group ISC was significantly higher than between-group ISC. Visual fixation patterns were significantly more similar among participants who made consistent stability judgments compared to those who made opposing judgments ($t(54) = 2.89$, $p = 0.005$), supporting the idea that shared perceptual strategies underlie similar intuitive physical judgments.

Discussion

Our study demonstrates that CNNs trained on large-scale visual datasets to perform stability judgment tasks can better predict human stability judgments than physics engine-based models. This finding aligns with prior research on CNNs as effective models of human visual processing (Yamins & DiCarlo, 2016; Kriegeskorte, 2015; Lindsay, 2021). Gaze data further revealed that both humans and the CNN attended to similar visual features when making stability judgments. Notably, despite being trained on ground-truth outputs from a physics engine, the CNN's predictions more closely matched human judgments than the ground-truth itself.

The divergence between human and physics-engine predictions suggests that intuitive physical judgments may rely on learned heuristics derived from visual patterns rather than explicit physical computations. CNNs, lacking any innate understanding of physics, learned to predict stability from visual patterns alone. The model's superior performance in predicting human responses compared to physics-engine-based ratings suggests that human 'errors' may in fact reflect adaptive strategies that are better suited for interpreting physical stability in everyday contexts. For example, people might judge an asymmetrical tower with a jutting block as unstable- this judgment reflects the statistical regularity that such structures often collapse, even if that particular tower was stable.

Further supporting this notion, CNN-generated importance maps corresponded strongly with human fixation patterns, suggesting that both rely on similar visual cues to make stability judgments. Our occlusion study confirmed that regions critical for the CNN's stability predictions closely aligned with areas humans fixated on, emphasizing the shared reliance on key visual features. This correlation underscores the utility of CNNs as computational models for studying perceptual heuristics in intuitive physics tasks.

These findings may contribute to hybrid cognitive models that propose that intuitive physics judgments integrate simulation-based reasoning with heuristics (Smith, Battaglia, & Tenenbaum, 2023). While mental simulation may play a role in complex or unfamiliar scenarios, our results suggest that humans can adopt faster, heuristic-driven strategies, particularly when time or information is limited. This dual-

process perspective aligns with theories of decision-making, where fast, experience-based processes complement slower, deliberative reasoning (Evans & Stanovich, 2013).

Future work could directly compare CNNs with simulation-based cognitive models, such as the Intuitive Physics Engine (IPE; Battaglia et al., 2013), to evaluate which framework more closely approximates human intuitive judgments. While the IPE has been shown to capture human error patterns (Zhang et al., 2016; Conwell et al., 2019), our focus here was to use a physics engine as an objective benchmark to assess divergence from ground truth. This approach enables clearer interpretation of whether CNNs and humans share systematic deviations from physical reality.

Our study thus highlights the value of using CNNs to explore how visual heuristics may emerge from statistical regularities in the environment, offering a data-driven approach to modeling intuitive physical reasoning. By probing the CNN's learned features, researchers can identify statistical regularities that underpin human judgments, providing a data-driven approach to studying heuristics. For instance, heuristics like symmetry or balance likely emerge from their correspondence to statistical patterns in natural environments. CNNs, as interpretable model systems that can be perturbed, manipulated, and probed at scale, allow us to investigate whether these features align with the heuristics humans use during intuitive physics tasks (Zhang, Wu, & Zhu, 2018).

Additionally, our findings emphasize the role of visual attention in shaping physical inference. Participants who focused on similar regions of an image tended to make the same judgments, as reflected in the high inter-subject correlation of gaze patterns. This suggests that shared attentional strategies contribute to agreement in stability judgments. Conversely, variability in gaze patterns across participants may underlie differences in their intuitive physics judgments, highlighting how variations in how visual information is sampled can lead to divergent conclusions. For example, participants might focus on blocks that appear to compromise the stability of a tower before judging it as unstable or, alternatively, look at parts of the tower that appear to counterbalance these blocks before judging it as stable. These differences may arise from individual biases in prioritizing specific features of the scene, such as asymmetries or areas of potential instability.

Building on this, the model's importance maps provide a valuable framework for identifying "stability" and "instability points" within images, for example, by examining aspects of an image that increase or decrease model estimates of stability. These points allow us to understand why participants prioritize certain visual features when making intuitive physical judgments and provide a structured approach for investigating how visual attention influences physical reasoning.

Additionally, exploring how external motivations may bias attention to stability and instability points could provide new insights into the interaction between goals, attention, and judgment. For example, financial incentives might cause individuals to prioritize instability points when classifying a tower as unstable, leading to systematic shifts in gaze patterns. This idea is supported by previous research suggesting that people are more likely to interpret ambiguous visual information in ways that align with their motivations (Calabro, Lyu, & Leong, 2023). Investigating such motivational influences could deepen our understanding of the cognitive mechanisms underlying intuitive physical reasoning.

Finally, while our study focused on controlled 2D block towers, real-world physical reasoning involves more dynamic and complex environments. Future research should extend these methods to naturalistic settings to assess whether reliance on visual heuristics generalizes. These extensions would enhance the ecological validity of CNN-based models and provide further insights into the interplay between visual attention, motivation, and intuitive physics in real-world contexts.

Conclusions

Our findings have important implications for cognitive science, particularly in understanding how humans perceive and reason about the physical world. By demonstrating that CNNs can predict human eye-gaze patterns and intuitive judgments about physical stability, our study contributes to the extensive body of literature suggesting that human perceptual judgments are supported by statistical learning. The observed parallels between CNNs and human perception suggest that similar underlying computational principles may govern both systems. By leveraging CNNs as models of human perception, we gain a tool to explore these predictive mechanisms in greater detail, providing a bridge between empirical studies of attention and theoretical accounts of intuitive physics.

Our study contributes to the cognitive science literature by providing a computational model that aligns closely with human perception and attention in intuitive physics tasks. This work sets up future research that could examine how intuitive physics judgments vary with development, individual differences, and experience, as well as how external factors like motivation, goals, or context influence these judgments. By bridging the gap between statistical learning and cognitive heuristics, we hope to deepen our understanding of the mechanisms underlying human perception and decision-making in physical environments. More broadly, our findings support the view that intuitive physical reasoning may reflect adaptive, experience-based strategies that are computationally efficient and ecologically rational. CNNs offer a scalable framework for formalizing and testing these strategies across tasks and domains.

References

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.

Borji, A., & Itti, L. (2013). State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 35*(1), 185–207.

Calabro, R., Lyu, Y., & Leong, YC. (2023). Trial-by-trial fluctuations in amygdala activity track motivational enhancement of desirable sensory evidence during perceptual decision-making. *Cerebral Cortex, 33*(9), 5690-5703.

Callaway, F., Hamrick, J.B., & Griffiths, T.L. (2017). Discovering simple heuristics from mental simulation. *Proceedings of the 39th Annual Meeting of the Cognitive Science Society: Computational Foundations of Cognition,* 1703-1708.

Conwell, C., Doshi, F., & Alvarez, G.A. (2019). Human-like judgments of stability emerge from purely perceptual features: evidence from supervised and unsupervised deep neural networks. *Conference on Cognitive Computational Neuroscience,* 605-608.

Dosovitskiy, A., Fischer, P., Springenberg, J.T., Riedmiller, M., & Brox, T. (2016). Discriminative Unsupervised Feature Learning with Exemplar Convolutional Neural Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 38(9), 1734-1747.

Evans, J. S. B. T., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science, 8*(3), 223–241.

Groth, O., Fuchs, F.B., Posner, I., & Vedaldi, A. (2018). ShapeStacks: learning vision-based physical intuition for generalised object stacking. *Proceedings of the European Conference on Computer Vision*, 684–701.

Hayhoe, M., & Ballard, D. (2005). Eye movements in natural behavior. *Trends in Cognitive Sciences, 9*(4), 188–194.

Kelley, L.A. & Kelley, J.L. (2014). Animal visual illusion and confusion: the importance of a perceptual perspective. *Behavioral Ecology, 25*(3), 450-463.

Kriegeskorte, N. (2015). Deep neural networks: A new framework for modeling biological vision and brain information processing. *Annual Review of Vision Science, 1*, 417–446.

Kümmerer, M., Wallis, T. S. A., & Bethge, M. (2015). Deep Gaze I: Boosting Saliency Prediction with Feature Maps Trained on ImageNet. *International Conference on Learning Representations (ICLR).*

Lerer, A., Gross, S., & Fergus, R. (2016). Learning physical intuition of block towers by example. *Proceedings of Machine Learning Research.*

Lindsay, G.W. (2021). Convolutional neural networks as a model of the visual system: past, present, and future. *Journal of Cognitive Neuroscience, 33*(10), 2017-2031.

Liu, Y., Ayzenberg, V., & Lourenco, S.F. (2024). Object stability serves humans' intuitive physics of stability. *Scientific Reports, 14, 1701.*

Ludwin-Peery, E., Bramley, N.R., Davis, E., & Gureckis, T.M. (2021). Limits on simulation approaches in intuitive physics. *Cognitive Psychology, 127*(2021), 101396.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why Should I Trust You? Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD).* 1135–1144.

Sanborn, A.N., Mansinghka, V.K., & Griffiths, T.L. (2013). Reconciling intuitive physics and Newtonian mechanics for colliding objects. *Psychological Review, 120*(2), 411-437.

Smith, K.A., Battaglia, P.W., & Tenenbaum, J.B. (2023). Integrating heuristic and simulation-based reasoning in intuitive physics. *PsyArXiv.*

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: heuristics and biases. *Science, 185*(4157), 1124-1131.

Wu, J., Yildirim, J., Lim, J.J., Freeman, B., & Tenenbaum, J. (2015). Galileo: perceiving physical object properties by integrating a physics engine with deep learning. *Advances in Neural Processing Systems 28*, 7-12.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R., & Bengio, Y. (2015). Show, attend and tell: neural image caption generation with visual attention. *Proceedings of the International Conference on Machine Learning*, *37*, 2048-2057.

Yamins, D.L.K., & DiCarlo, J.J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience 19*(3), 356-365.

Zhang, Q., Wu, Y.N., & Zhu, SC. (2018). Interpretable Convolutional Neural Networks. *arXiv.*

Zhang, R., Wu, J., Zhang, C., Freeman, W.T., & Tenenbaum, J.B. (2016). A comparative evaluation of approximate probabilistic simulation and deep neural networks as accounts of human physical scene understanding. *Proceedings of the Annual Meeting of the Cognitive Science Society, 38.*