Humans and deep neural networks can use perceptual features to determine object stability

Lauren S. Aulet (laulet@andrew.cmu.edu)
Department of Psychology, 5000 Forbes Ave
Pittsburgh, PA 15206 USA

Evren M. Konuk (emkonuk@andrew.cmu.edu)
Department of Psychology, 5000 Forbes Ave
Pittsburgh, PA 15206 USA

Jessica F. Cantlon (jcantlon@andrew.cmu.edu)
Department of Psychology, 5000 Forbes Ave
Pittsburgh, PA 15206 USA

Abstract:

There is extensive debate about whether judgments of physical stability (e.g., whether a stack of blocks will fall) rely on low-level perceptual features or mental simulation. In the present work, we evaluated whether deep neural networks (DNNs) trained on ImageNet, which are thought to rely only on low-level image features (and cannot perform simulations) can discriminate images of stable and unstable block towers. Moreover, we evaluated whether human adults were affected by the stability of a distractor block tower image (i.e., same or different stability category) when performing an exact match-to-sample task. We found that DNNs discriminated stable and unstable block towers significantly above chance, and did so across a variety of stimulus perturbations. Furthermore, we found that human participants were significantly influenced by the stability of the block tower images, even in a task where mental simulation was highly improbable. Taken together, these results suggest there are visual features that are diagnostic of physical stability and are 'perceived' by both DNNs and humans.

Keywords: intuitive physics; visual perception; deep neural networks; human cognition; physical reasoning; mental simulation

Introduction

In the field of cognitive science, the most common theory of physical reasoning is that of mental simulation (Hegarty, 2004; Johnson-Laird, 2002; Kubricht, Holyoak, & Liu, 2017). These theories typically posit that humans use a probabilistic simulation to predict outcomes of physical scenarios (i.e., a "physics engine" in the mind; Battaglia, Hamrick & Tenenbaum, 2013).

Others have criticized *probabilistic mental simulation theory* (PMST) on both theoretical and empirical grounds (Davis & Marcus, 2016; Ludwin-Peery, Bramley, Davis & Gureckis, 2021). Opponents of PMST typically posit that visual features, alone, are sufficient for making a variety of physical judgments (Liu, Ayzenberg, & Lourenco, 2024). For example, Conwell and colleagues (2019) showed that deep neural networks (DNNs; e.g., ResNet18), trained on ImageNet, can successfully discriminate stable and unstable block towers. Furthermore, the variability in accuracy across block tower images was significantly correlated between DNNs and humans, suggesting possible similarity in the mechanisms underlying DNN and human performance.

However, there remain open questions about whether DNNs' ability to discriminate stable and unstable block tower images can generalize to other stimuli or task contexts (Ullman, Spelke, Battaglia, & Tenenbaum, 2017). If DNNs cannot discriminate physical stability across stability-irrelevant stimulus changes, then this would suggest that the image features may not be sufficient to explain human performance, and other processes (i.e., mental simulation) may be needed to generalize across diverse or novel contexts.

Model Performance

First, we asked whether DNNs can discriminate stable and unstable block tower images across a set of stimulus changes (color change, size change, viewpoint change, scrambled, line drawing). Capacity to discriminate block tower stability with above chance

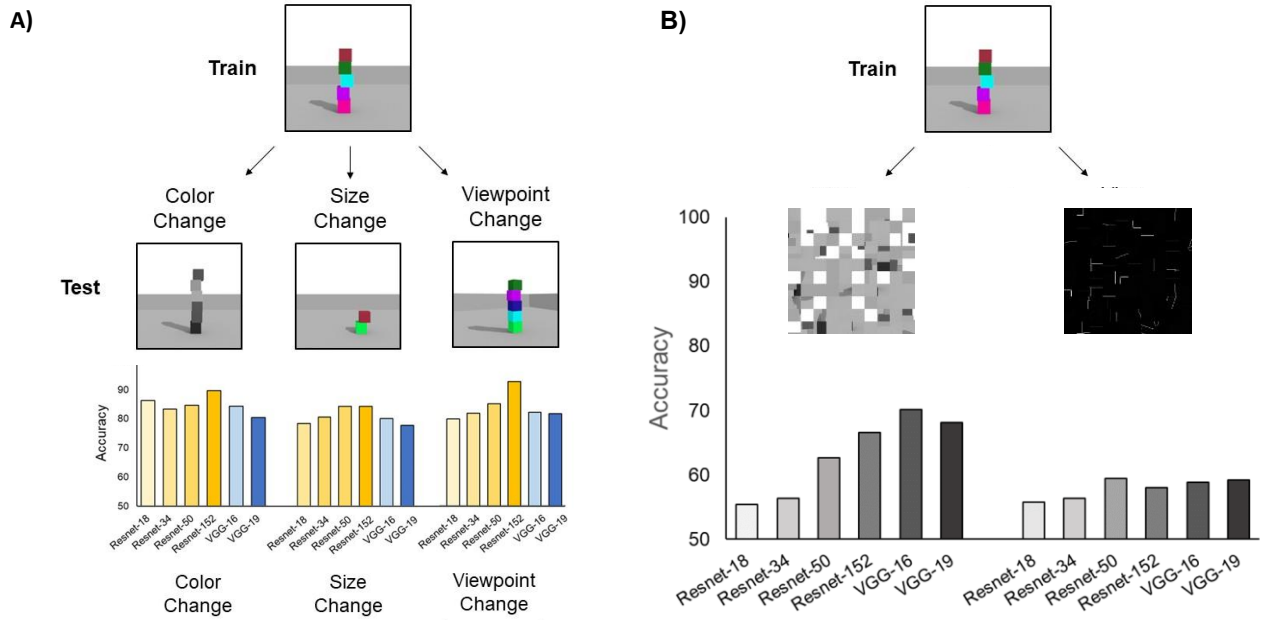

Figure 1: DNN accuracy at discriminating stable and unstable block towers across a variety of stimulus manipulations.

accuracy would suggest the presence of low-level image features that are sufficient for determining block tower stability.

Methods

All DNNs used (ResNet18, ResNet34, ResNet50, ResNet152, VGG16, and VGG19) were pretrained on ImageNet. All models were then trained on 8,000 color images of block towers (see Fig 1). For details regarding stimulus creation and the determination of ground truth stability of the block tower images, see Conwell et al., 2019.

Results

We found that all DNNs evaluated were significantly above chance (.50) at discriminating block tower stability at test, even when test stimuli varied systematically from the training stimuli (i.e., color versus black-and-white, different number of blocks, different image viewpoint; see Fig 1A). Interestingly, the DNNs remained significantly above chance even when test stimuli varied dramatically (i.e., 10x10 scrambled, or line drawings; see Fig 1B).

Human Performance

Second, we asked whether adult humans are sensitive to low-level image features that distinguish between stable and unstable block towers. Critically, we were interested in evaluating whether human subjects are sensitive to block tower stability in the absence of any mental simulation. Accordingly, we used a task that did not require any explicit judgment of object stability (i.e., identical match-to-sample). To rule out the possibility that the stimuli can trigger automatic mental simulation of physical stability (Solomon & Barsalou, 2004; Stanfield & Zwaan, 2001), even in the absence of explicit stability judgments, we presented all block tower stimuli in an inverted orientation (i.e., rotated 180 degrees).

Methods

The match-to-sample task was created in PsychoPy (Peirce, 2019). On each trial of this task, a 'sample' stimulus was presented in the center of the screen for 500ms. Following a 1s interstimulus interval, two stimuli were presented, on the left and right sides of the screen. Participant were instructed to press the 'q' and 'p' keys to indicate whether the left or right stimulus, respectively, was identical to the previous sample stimulus shown. Participants were adult humans (N = 10) who received course credit for their participation. All procedures were approved by the local Institutional Review Board.

Results

Accuracy on the match-to-sample task (mean accuracy = .93) was significantly above chance (.50), $p < .001$. Only correct trials were analyzed further. For each participant, we calculated the mean reaction time (RT) in milliseconds, for trials in which the incorrect (i.e., non-matching) stimulus was from the same stability category as the sample image ('Same' trials; i.e., both stimuli were stable, or both stimuli were unstable) and for trials in which it was from a different stability category (i.e., one stimulus was stable and the other was unstable). We found that mean RTs were significantly larger for 'Same' trials than for 'Different' trials. In other words, participants were slower at correctly identifying the sample image when the distractor image was more similar in stability. This finding suggests that participants were sensitive to perceptual features that were diagnostic of physical stability, even in the absence of an explicit physical stability judgment.

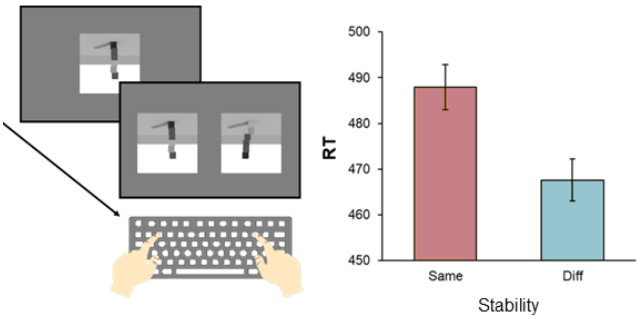

Figure 2: A depiction of the match-to-sample task (left) and mean RT (ms) for trials in which the distractor image was from the same or different stability category (right).

Discussion

Our results demonstrate that both DNNs and humans are sensitive to image-level features that are diagnostic of physical stability. Moreover, these perceptual features affect human performance even when physical stability is irrelevant for the task. These findings suggest the information necessary for successful stability judgments is present in the absence of mental simulation.

In future work, it will be necessary to determine in what contexts humans use perceptual features versus mental simulation to guide their judgments about the physical world. For example, it is unclear how computationally costly mental simulation is, and, accordingly, whether a mechanism like mental simulation is a plausible explanation of physical reasoning early in infancy, and in non-human animals (Chiandetti & Vallortigara, 2011; Needham & Baillargeon, 2003).

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*, 18327-18332.

Chiandetti, C., & Vallortigara, G. (2011). Intuitive physical reasoning about occluded objects by inexperienced chicks. *Proceedings of the Royal Society B: Biological Sciences*, *278*, 2621-2627.

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

Davis, E., & Marcus, G. (2016). The scope and limits of simulation in automated reasoning. *Artificial Intelligence*, *233*, 60-72.

Fischer, J., & Mahon, B. Z. (2021). What tool representation, intuitive physics, and action have in common: The brain's first-person physics engine. *Cognitive Neuropsychology*, *38*, 455-467.

Hegarty, M. (2004). Mechanical reasoning by mental simulation. *Trends in Cognitive Sciences*, *8*, 280-285.

Johnson-Laird, P. N. (2010). Mental models and human reasoning. *Proceedings of the National Academy of Sciences*, *107*, 18243-18250.

Kubricht, J. R., Holyoak, K. J., & Lu, H. (2017). Intuitive physics: Current research and controversies. *Trends in Cognitive Sciences*, 21, 749-759.

Li, W., Azimi, S., Leonardis, A., & Fritz, M. (2016). To fall or not to fall: A visual approach to physical stability prediction. *ArXiv*.

Liu, Y., Ayzenberg, V., & Lourenco, S. F. (2024). Object geometry 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*, 101396.

Needham, A., & Baillargeon, R. (1993). Intuitions about support in 4.5-month-old infants. *Cognition*, *47*, 121-148.

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., ... & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, *51*, 195-203.

Solomon, K. O., & Barsalou, L. W. (2004). Perceptual simulation in property verification. *Memory & Cognition*, *32*, 244-259.

Stanfield, R. A., & Zwaan, R. A. (2001). The effect of implied orientation derived from verbal context on picture recognition. *Psychological Science*, *12*, 153-156.

Ullman, T. D., Spelke, E., Battaglia, P., & Tenenbaum, J. B. (2017). Mind games: Game engines as an architecture for intuitive physics. *Trends in Cognitive Sciences*, *21*, 649-665.