Shared Representations of Stability in Humans, Supervised, & Unsupervised Neural Networks

Colin Conwell
Department of Psychology
Harvard University
Cambridge, MA 02139
conwell@g.harvard.edu

Fenil Doshi
Department of Psychology
Harvard University
Cambridge, MA 02139
fenil_doshi@fas.harvard.edu

George A. Alvarez
Department of Psychology
Harvard University
Cambridge, MA 02139
alvarez@wjh.harvard.edu

Abstract

At a glance, the human visual system transforms complex retinal images into generic feature representations useful for guiding a wide range of behaviors. Here, we provide evidence that the feature representations embedded in purely feedforward neural networks are sufficient to explain seemingly high-level human judgments: in this case, the stability of randomly arranged block towers. To do this, we first show that we can linearly decode stability from the features of two deep neural networks – a supervised network trained on ImageNet, and a variational autoencoder trained only to reconstruct images of block towers from various perspectives – neither of which were ever taught stability per se. We then derive a set of image-computable features to use as predictors of performance, finding that the stability judgments of both human subjects and the neural net decoders are best predicted by the same feature. Our findings suggest overall that at least some aspects of seemingly higher-level reasoning in a now paradigmatic intuitive physics task may be grounded in direct readouts of purely perceptual features.

1 Introduction

In a fraction of a second, humans can extract surprisingly sophisticated information: the trustworthiness of a face, the centroid of complex shapes, and – directly relevant to the present work – even the stability of a tower of randomly arranged blocks [4]. In some cases, this extraction takes less than a 20th of a second. The speed at which we can decode these bits of meaning in general is taken as evidence that the decoding happens not at the level of some abstract cognitive process, but directly in the rapid cascade of perceptual processing that occurs immediately after the presentation of a stimulus. This 'feedforward' processing – predominantly encapsulated from computations performed elsewhere in the brain – we now increasingly model using 'feedforward' deep neural networks. These networks, while diverse in form and function, are by and large trained on some variant of a singular class of task: the nonlinear regression of raw inputs (pixels, waveforms, words and numbers) onto various predictors. This class of task goes by many names, depending on the context and community in which it's used, but perhaps most relevant to human behavior is the name *statistical pattern recognition*.

A growing body of evidence suggests that the pattern recognition done by deep neural networks is a superlative model of the pattern recognition done in biological cortex, and that the representations

1st Workshop on Shared Visual Representations in Human and Machine Intelligence (SVRHM), NeurIPS 2019.

these networks learn tend to correspond strikingly well with biological reality [1, 16, 13]. But are these networks useful as models of judgments that extend beyond the typical purview of sense-percepts? How do they fare in purportedly more complex domains, such as intuitive physical reasoning?

Given the complexity and latent structure of the physical world, predominant models of intuitive physics posit that our intuitive physical capabilities are at their core the product of a cognitive architecture that includes a more or less complete 'physics engine', akin to the kind deployed in video games and computer graphical animation [2, 15]. Many of these models are innately equipped with effectively all the parameters necessary to perform complex simulations of physical scenarios in real time. Inference in this formulation is accomplished by iteratively and repeatedly sampling a simulator – primed by perception but powered (sometimes exclusively) by cognition.

In the current report, we explore an alternative model of intuitive physics based on the pattern recognition capabilities of deep neural networks, wherein physical inference is reformulated as a problem of identifying those perceptual features that serve as optimal proxies for the real physical properties that produce them. Previous work [3, 11, 17] has mainly focused on training networks end to end in a fully supervised fashion, developing features directly for explicitly physical targets. Here, we explore another possibility: that features learned by deep neural networks trained for other tasks may nevertheless encode physically relevant properties that serve as the basis for physical inference.

2 Methods

To test this hypothesis, we fit linear classifiers of stability on the learned features of two model classes: a supervised neural network trained only on image recognition with ImageNet, and an unsupervised neural network trained only to reconstruct images of block towers (never with provision of the groundtruth stability). We then compare the responses of these classifiers to those of human observers using a set of stimulus-computable features meant to disambiguate not only whether neural net decoders can adequately gauge stability, but how.

2.1 Stimulus Set

Adapting a technique specified by [17], we generated an image dataset of stacked blocks, all of the same size ($1m^3$), with enough horizontal jitter in each block's position that towers had a 50/50 chance of falling. We varied the number of blocks from 2-6. The groundtruth for whether a tower will fall can be determined by computing at each junction of blocks the mean position (centroid) of all the blocks above the junction and comparing it to the centroid of the block beneath. If the centroid of the blocks above extends beyond the edge of the block beneath (at any junction), the tower should fall. (We confirmed this groundtruth calculation with forward simulations of the animator producing our images). In one of two datasets we generated with this method (called 'Perspective'), we allow some variance in the camera. In the other (called 'Direct'), we situate the camera directly in front of the blocks, with the camera focused directly at the tower's center.

2.2 Behavioral Tasks

Human subjects (N = 81, from Amazon Mechanical Turk) were shown a series of towers (100 stable, 100 unstable, 200 total, randomized in order of presentation) and given a binary forced choice task, designating each tower as stable or unstable. In the conditions relevant to this analysis, we put no time limit on the presentation of the stimuli, nor did we constrain response times. Tower sizes varied across subjects, but each subject rated only one size.

2.3 Models & Modeling Tasks

For our supervised image recognition network, we used Resnet18 [6] pre-trained on ImageNet as a fixed feature extractor, freezing all the layers of the network except the batch normalization layers – a technique that maintains the integrity of the features learned by the convolutional and nonlinear filters of the network, but accounts for vacillations in the statistics of the image set currently being processed [7]. For our unsupervised image reconstruction network, we used a variational autoencoder [8] with a randomly initialized Resnet18 [6] as the encoder and a latent space of 128 dimensions, trained on the full range of block tower sizes in the 'Perspective' dataset (roughly 5000 images per

tower size, 25,000 images total) using a mean squared error reconstruction loss and a generative adversarial loss function [12, 5] in place of the standard Küllback-Liebler divergence, allowing the model to 'learn' the variational prior (a Gaussian) imposed on it. Importantly, and in contrast to other approaches that attempt to disentangle certain properties in the latent space using techniques like minibatch discrimination [9], we leave the latent space of our autoencoder fully entangled.

For both our image recognition and image reconstruction models, we decoded stability from the outputs of the last convolutional layer using a multilayer linear perceptron, trained on the 'Direct' dataset with Adam optimization and a cyclical learning rate deduced from search [14]. We used the last convolutional layer of these networks (rather than the fully connected layer or 128-dimensional latent space) to maintain architectural consistency across our feature extractors. For any given size of tower, we held the process of feature extraction constant, but varied the process of linear decoding, always training and testing the classifier on the same size of tower. Classifiers fit on both the image recognition network and the image reconstruction network were trained using features from 25,000 towers per tower size and tested on the benchmark towers of the same size rated by human subjects (200 images per tower size; 1000 total). Crucially, we reserved the set of images from the human behavioral studies as a test set. Neither model ever trained on these images prior to evaluation. We reinitialized and fit classifiers 5 times per tower size, resulting in a total of 50 classifiers (25 for the image recognition model and 25 for the image reconstruction model). We treated each of these classifiers as individual subjects, comparable directly to our human subjects. All analyses were identical across human and machine.

2.4 Feature Analysis

For each stimulus in the test set, we considered 12 hypothetical features a human or machine might use to accomplish the task. The features differ in whether they emphasize local information, or statistical information aggregated over multiple local measurements. Importantly, they also differ in how well they approximate the groundtruth stability of a given tower. Each feature is quantifiable in the sense that it constitutes some property of the visual array and differs across exemplars, but contains no explicitly physical information (e.g. the mass of the blocks, or the force of gravity). Below are more detailed descriptions of each of the 12 features we used in our analysis.

- Configural deviation: the mean and max values for the distance of each block from the centroid of all the blocks above it. The max value of this feature is a perfect predictor of groundtruth in that any value beyond a given threshold (half the width of a block) means the tower will fall. Any value below that threshold means the tower will remain upright.
- Local (pairwise) deviation: the mean and max values for the distance of each block in the tower from the block above it, irrespective of other blocks.
- Global deviation: the mean and max values for the distance of each block from the overall centroid of the tower (the centroid of all the blocks considered together).
- Number of instabilities: the number of junctions in the tower shown by groundtruth calculations to be unstable. This feature is also a perfect predictor of groundtruth, in that any value of 0 means the tower is stable and any value above 0 means the tower is unstable.
- Percent unstable: the number of unstable junctions in the tower as a proportion of the total number of junctions. Like number of instabilities (and for the same reasons) this feature is another perfect predictor of groundtruth.
- Horizontal extent: The horizontal distance from the right edge of the rightmost block in the tower to the left edge of the leftmost block in the tower: the tower's width.
- Vertical extent: The vertical distance from the bottom edge of the bottommost block to the upper edge of the uppermost block: the tower's height.
- Alignment distance: the numerically determined minimum distance each block must be moved to return the tower to a perfectly stable configuration, wherein each block is perfectly aligned with the others.
- Minimum distance to stability: the minimum each block must be moved to return the tower to a minimally stable configuration, wherein there are no unstable junctions. Because a value of 0 means the tower is already stable, and any value above 0 means the tower is unstable, this feature is a yet another perfect predictor of groundtruth.

To determine which of our hypothesized features was most predictive of performance, we used a combination of penalized regressions, random forest variable importance analysis and receiver operating characteristic curves. In the case of penalized regression, we simultaneously fit three types – ridge, lasso and single-component partial least squares – to assess the reliability of each feature as a predictor of performance and to mitigate what was often the high multicollinearity of the features considered together. Each regression was fit to predict the response of a given set of subjects (either human or machine) to each image using as predictors the full suite of feature values for that image. In the case of the lasso and ridge regression, we designated the winning feature as the feature with the largest coefficient following the cross-validated regularization procedure. In the case of partial least squares, we considered both the largest coefficient and the largest influence on the projection (effectively equivalent to the largest loading on the first component in a principal component analysis). For the random forest variable importance analysis, we once again required the model to predict the individual responses of a given subject type using the full range of features for each image, arbitrating the most predictive feature by its mean decrease in accuracy and mean decrease in Gini coefficient (a measure of how much a given variable streamlines the decision tree). Finally, for the receiver operating characteristic curves, we fit a series of mixed effects logistic regressions individually for each feature, once again predicting individual responses to each image. The logistic regression with the highest area under the curve we designated as the most predictive feature of performance.

While these analyses often converged on the same feature, this was not always the case – a byproduct to some extent of the multicollinearity inherent to the feature space. Where the analyses diverged, we designated the most predictive feature as the feature that ranked highest in the majority of analyses with an overall score marked as the number of metrics in agreement divided by the total number of metrics – a total of 7 (4 penalized regression measures, 2 random forest variable importance measures, 1 area under the curve).

3 Results

3.1 Overall Performance

Human performance was generally high for the full range of blocks in the range we tested (with roughly 84% accuracy on average). The performance of the linear classifiers trained on features from the ImageNet-pretrained Resnet18 (henceforth 'Resnet18-ImageNet') and the Resnet18 embedded in the autoencoder (henceforth 'Resnet18-Autoencoder) both produced slightly higher performance (with 87.7% and 86% average classification accuracy, respectively). The higher performance of the models seems almost certainly to be a product of the larger quantity and diversity of examples in the training set, a number we could theoretically reduce to more closely approximate human performance – albeit with some caveats (see Appendix). A linear regression of accuracy on tower size unveiled a slight, but significant negative slope for both humans ($b = -0.024$, $p < 0.001$) and machines ($b = -0.03$, $p < 0.001$ & $b = -0.046$, $p < 0.001$ for Resnet18-ImageNet & Resnet18-Autoencoder, respectively) – suggesting that an increased difficulty of classification with larger and larger tower sizes was a trend shared across all the subject types we tested (Fig. A.1 in Appendix).

3.2 Feature Analysis

The feature most predictive of human performance was the maximum configural deviation – the maximum distance of any block from all the blocks above it, ranked highest in 6 out of 7 metrics (with the mean configural deviation providing the largest decrease in accuracy for the random forest variable importance analysis). This feature alone explains 91.6% of the variance in human responses.

The feature most predictive of Resnet18-ImageNet and Resnet18-Autoencoder's performance's was also the maximum configural deviation, ranked highest in 6 out of 7 metrics for both models (with the maximum *global* deviation – the maximum distance of each block from the overall centroid of the tower – proving slightly more predictive in terms of the random forest's mean decrease in accuracy; see Appendix D & E). The maximum configural deviation explains 94.7% of the variance in the responses of of Resnet18-ImageNet and 93.9% of the variance of Resnet18-Autoencoder (Fig. 1)

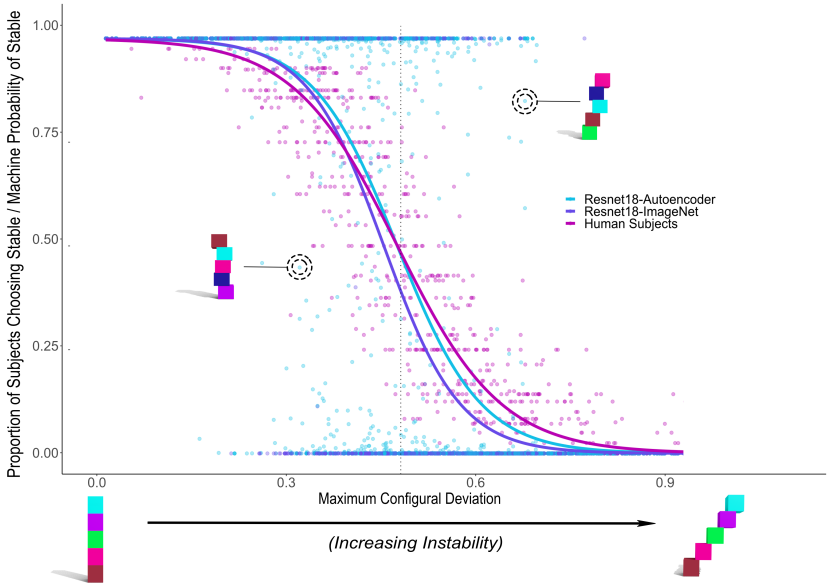

Figure 1: Psychophysical curves of human and machine performance in the block towers task predicted by the maximum configural deviation. On the x axis is the maximum configural deviation, ranging from perfectly stable at the lower end to very unstable at the upper end. On the y axis is the machine's confidence and the proportion of subjects choosing 'stable'.

4 Discussion

The success of our linear classifiers in decoding the stability of block towers from the features of supervised and unsupervised deep neural networks is not evidence that they have learned the same representations present in human perceptual systems: decades of cognitive science have shown those representations to be more rich, more flexible and more robust than the representations we have explored here. What the success of our linear classifiers does mean is that there exists some linear mapping between the purely perceptual representations learned by a deep neural network and the representations powering the inferences of human subjects in a task traditionally conceptualized as requiring a heavy dose of higher-order abstraction. While this work does not arbitrate on the capacity for such abstraction, it does suggest we may not always need it – and that statistical shortcuts via perceptual features may well trump fully fledged simulation in the pinch of computational pressure. All this to say, we may not always need physics to make physical inferences.

Future work will attempt to further complete the cartography of correspondence between human and machine by pushing and plying how we learn the representations we do, and probing why, despite immense divergences in the material substrate on which these algorithms are instantiated, the correspondences persist. The autoencoder we have included here – though in many ways an undercomplete example precisely because of its highly constrained, synthetic input space – is a nod to the necessity of rethinking how our perceptual systems are tuned, and what features they might develop in the process of the tuning. The more kaleidoscopic our representational palette, the more robust it is to the uncertainties and perturbations we invariably encounter, and the more conducive to a properly calibrated response.

5 Acknowledgements

Many thanks to Tim Menke for assistance in calculating the features.

References

[1] P. Bashivan, K. Kar, and J. J. DiCarlo. Neural population control via deep image synthesis. *Science*, 364(6439):eaav9436, 2019.

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

[3] C. Conwell and G. Alvarez. Modeling the intuitive physics of stability judgments using deep hierarchical convolutional neural networks. In *2018 Conference on Cognitive Computational Neuroscience*. Cognitive Computational Neuroscience, 2018.

[4] C. Firestone and B. Scholl. Seeing stability: Intuitive physics automatically guides selective attention. *Journal of Vision*, 16(12):689–689, 2016.

[5] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[6] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[7] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[8] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[9] T. D. Kulkarni, W. F. Whitney, P. Kohli, and J. Tenenbaum. Deep convolutional inverse graphics network. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 2539–2547. Curran Associates, Inc., 2015.

[10] B. Lakshminarayanan, A. Pritzel, and C. Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, pages 6402–6413, 2017.

[11] A. Lerer, S. Gross, and R. Fergus. Learning physical intuition of block towers by example. *arXiv preprint arXiv:1603.01312*, 2016.

[12] A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow, and B. Frey. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*, 2015.

[13] T. Serre, A. Oliva, and T. Poggio. A feedforward architecture accounts for rapid categorization. *Proceedings of the national academy of sciences*, 104(15):6424–6429, 2007.

[14] L. N. Smith. Cyclical learning rates for training neural networks. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 464–472. IEEE, 2017.

[15] T. D. Ullman, E. Spelke, P. Battaglia, and J. B. Tenenbaum. Mind games: Game engines as an architecture for intuitive physics. *Trends in Cognitive Sciences*, 21(9):649–665, 2017.

[16] D. L. Yamins, H. Hong, C. F. Cadieu, E. A. Solomon, D. Seibert, and J. J. DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624, 2014.

[17] R. Zhang, J. Wu, C. Zhang, W. T. Freeman, and J. B. Tenenbaum. A comparative evaluation of approximate probabilistic simulation and deep neural networks as accounts of human physical scene understanding. *arXiv preprint arXiv:1605.01138*, 2016.

Appendix

A What of the 'Superhuman' Performance?

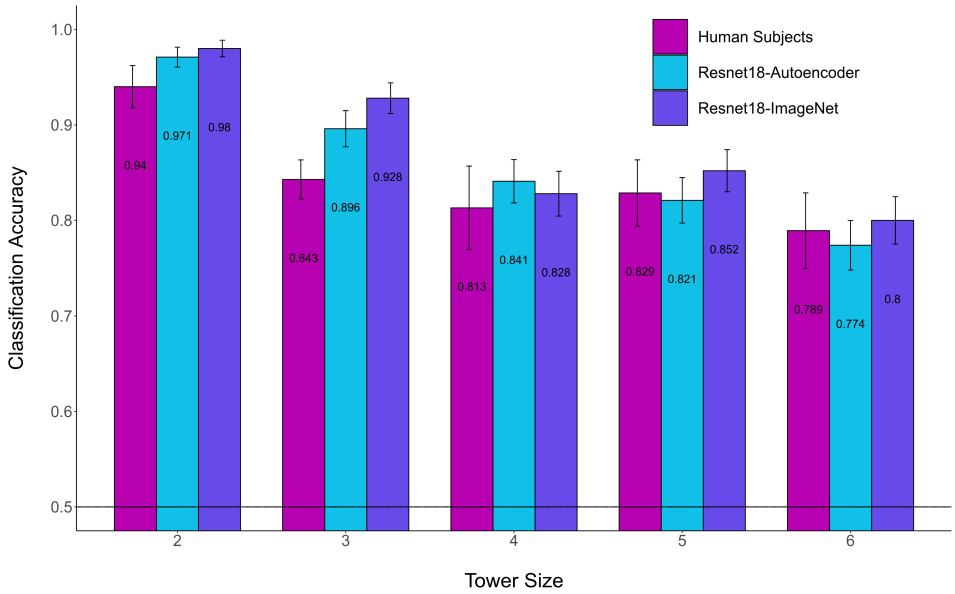

Figure A.1: Overall performance of humans and machines on the block towers test set. Error bare are 95% confidence intervals. Chance accuracy is 50%.

The superhuman performance of our classifiers (Fig. A.1) is in some ways one of the more striking bits of evidence against the hypothesis that humans and machines are similar. We might temper this evidence with two considerations: one, in only about 60% of the classifiers we trained does performance extend beyond the upper bound of human ability (14 / 25 classifiers for the Resnet18-Autoencoder; 19 / 25 classifiers for the Resnet18-ImageNet) and statistically significant differences are observed overall in only in the first two sizes of tower; two, the representations formed by the networks are at the moment relatively noiseless. Human representations suffer a wide variety of noise that make estimates of properties critical to inference (like the centroid of the blocks) more difficult to ascertain, and once ascertained, to maintain. Rather than brutishly decreasing the overall performance by training them less, we could use our understanding of the noise in human estimates to engender harder trials for our fully trained machine, lowering the performance to human levels by adding the stuff of human error.

B Item-Level Analysis

Is an image that's difficult for human subjects also difficult for machines? To answer this question, we correlated the aggregate responses of human subjects to each individual image (a proxy measure of difficulty in which harder images should be marked by less consensus than easier images) with the softmax probabilities produced by the decoders on the same set of images. Because the behavioral task was binary forced choice, we can measure consensus in terms of subjects choosing 'stable' or subjects choosing 'unstable'. In this case, we targeted the choice of "stable", correlating the proportion of subjects labeling a given stimulus as 'stable' with the model's softmax probability of 'stable' – a number between 0 and 1 interpretable more or less as the model's confidence that the stimulus in question was in fact 'stable' (Fig. 1, though see [10] for further discussion). The logic here is that the lower the consensus among human subjects as to whether a given stimulus was stable, the lower the model's confidence should be in its prediction. Confirming this logic, the Pearson correlation of human agreement scores and Resnet18-ImageNet's softmax probabilities was markedly high ($r = 0.814$, $p < 2.2e\text{-}16$); the same correlation with Resnet18-Autoencoder's softmax probabilities was only slightly lower ($r = 0.8$, $p < 2.2e\text{-}16$). These correlation coefficients suggest a

significant degree of overlap between the images humans found difficult and the images the model found difficult, though we note these coefficients may be somewhat inflated by the high performance overall of both human and machine.

C Interpolations in the Latent Space of the Autoencoder

In addition to developing representations powerful enough to undergird the explicit decoding of stability by a linear classifier, the Resnet18-Autoencoder we trained seemed also to develop strong implicit representations of stability, as evidenced by interpolations in the model's latent space that produced smooth generative samples in the transition from an unstable tower to a perfectly stable tower – an 'idealized' tower entirely absent from the network's training set (Fig. C.1).

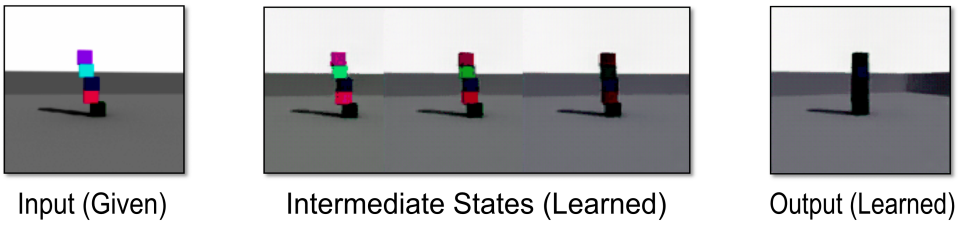

Figure C.1: Interpolations between an unstable tower and a perfectly stable tower in the latent space of a variational autoencoder never given stability labels. The image on the left was the only image not generated by the decoder.

These implicit representations may represent an inroad to exploring simulation not as abstract operations on graphically vectorized, probabilistic game engines, but as interpolation in learned feature spaces, where the distance between various representations can be leveraged to power more sophisticated types of inference beyond mere classification – inferences like the prediction of future physical states from current sensory input.

D On the Relationship Between Performance and Features

Given the generally high accuracy in stability judgments across humans and machines, is it inevitable that max configural deviation (a perfect predictor of groundtruth stability) would best predict performance? This might be the case were it not for the presence of three other features in our feature set that also perfectly predict stability – the number of instabilities, percent unstable and minimum distance to stability – as evident in their achieving an area under the curve of exactly 1.0 when used individually as predictors in logistic regressions predicting groundtruth itself (instead of individual choice, as in the main analysis). While the relative contributions of these features cannot be determined if accuracy is at 100%, our analysis suggests that gradations in performance below 100% are better predicted by max configural deviation than these other ground-truth indicators.

E Confirming the Most Predictive Feature of Model Performance through Targeted Disentanglement

To confirm that the maximum configural deviation was the most predictive feature of model performance not because of some statistical fluke, but for deeper structural reasons, we attempted to more systematically disentangle it from the second most predictive feature of model performance: the maximum global deviation (the maximum distance of each block from the overall centroid of the tower). To do so, we attempted (inasmuch as the variance our synthetic dataset allowed) to choose subportions of a held-out dataset wherein one of the two values was held constant while the other was left to vary. We succeeded in producing two new test sets, each controlling for one of our two features of interest. In the test set controlling for the maximum configural deviation, we achieved a relatively even balance of tower sizes and groundtruth stability by selecting blocks that varied

±0.57% of the range of maximum configural deviation values in the full dataset. (A more intuitive sense of this narrowed variance is to think of it in terms of the width of the blocks, in which case this value represents about ±0.75% the width of a block or $0.0075m^3$, arguably a perceptually negligible distance). In the test set controlling for the maximum global deviation, we achieved a similarly even balance of tower sizes and groundtruth stability by selecting blocks that varied ±0.17% of the range of maximum global deviation values (or about $\pm 0.0025m^3$, in terms of block width). We combined these test sets to create a new benchmark set of 400 images total (200 stable, 200 unstable, with between 25 and 50 images per tower size per groundtruth stability). We then evaluated each of our pretrained classifiers on this new benchmark set (recording responses only to the tower sizes on which they'd been trained). Given their similar profiles in the initial feature analysis, we combined the classifier responses from both of our feature extractors (Resnet18-ImageNet & Resnet18-Autoencoder) into one aggregate set of response profiles.

To gauge how predictive each of our two features was when controlling for the other, we calculated their areas under the curve in a series of logistic regressions regressing the choice of stable on the value of the target feature. The results of this test confirmed that that classification was more dependent on maximum configural deviation than maximum global deviation: the area under the curve was .81 for maximum configural deviation (holding global deviation constant), and .606 for the maximum global deviation (holding configural deviation constant). To confirm this difference, we conducted a bootstrapping analysis in which we resampled the responses of our classifiers (maintaining the same proportion of tower sizes and groundtruth stability across samples) and recorded the difference in area under the curve for each resampling. This analysis produced a mean bootstrapped difference of .175 (SD = .054). Only one of these thousand differences was negative (p = .001).

While controlling for one feature does not by any means guarantee that we've controlled for the rest, the results here do suggest that the penalized regressions and variable importance functions we described in the main analysis are plausible in their indication of the maximum configural deviation as the most predictive feature.

F On Generalization

The most common objection to the use of neural networks in the modeling of human cognition is their inability to 'generalize' outside of their training set. While this *might* be a valid criticism of neural networks more generally, when it comes to the modeling of human behavior more specifically, the criticism is somewhat less applicable. Each of the models we have trained here is meant to act as a standalone model of a given human behavior – in this case, stability judgments for a specific size of tower in a specially contrived artificial world — a self-contained umwelt. It is no different in this sense than a model fabricated by hand, in which the parameters set for a given experiment will rarely transfer unperturbed to an entirely new experiment. Even the intuitive physics engine designed by [17] was initialized with a host of parameters specific to the block towers task – including a 'poke' parameter that applied a lateralized force to the block towers during the course of simulation. We are not seeking to claim that the neural networks we have trained here are models of intuitive physics writ large, only that they capture a significant degree of explainable human performance in a paradigmatic intuitive physics task and have the major advantage of being image-computable.

It's worth clarifying here what generalization *could* mean in the context of fixed feature extraction, the method we've attempted here. Rather than creating a highly specific set of what might be called 'stability detectors' through end to end training ('stability detectors' a moniker we could confer only by more directly measuring the response profiles of individual artificial neurons, and even then only with some imagination), fixed feature extraction constrains our classifiers to figuratively 'work with what they've got'. 'What they've got' in the case of Resnet18-ImageNet especially is a set of features tailored for image recognition across a million, often highly divergent images. It is unlikely in the space of features learned by this model that there exist 'stability detectors' specific to our dataset. What is more likely is that there exist myriad feature combinations that higher-order inference can leverage to arbitrate various visual categories across a wide variety of inputs, and that some of these categories (like stability) may also be physical. In this sense, the combination of fixed feature extractor and linear classifier (often the formula for many of the 'transfer learning' tasks popular in modern machine learning) is in itself a test of the generalizability of features learned for natural image recognition to an almost entirely novel, synthetic domain.

Admittedly, this is just one of many possible interpretations of generalization. Generalization in the broadest sense is a decidedly ambiguous term, with numerous definitions that make it more or less plausible that human agents, as is often touted, generalize any more than the most massively over-fitted neural networks. The deepest ambiguity in the contest over generalization seems to be between interpolation within a domain of experience (e.g fitting a missing puzzle piece in a relatively complete board) and extrapolation outside that domain (e.g. predicting the contents beyond the edge of the puzzle), the former a perfectly reasonable behavior to expect of finite models, the latter decidedly less so. In future work, we intend to explore both the extent to which deep neural networks interpolate within their range of experience (e.g guessing the stability of 4 block towers when trained on 3 and 5) and the extent to which humans truly extrapolate outside *their* range of experience (e.g grappling with the physics of block towers made of 4D tesseracts instead of 3D cubes).