Measuring and predicting variation in the interestingness of physical structures

Cameron Holdaway*
Department of Psychology
UC San Diego
choldawa@ucsd.edu

Daniel M. Bear*
Department of Psychology
Stanford University
dbear@stanford.edu

Samaher F. Radwan
Department of Psychology
Stanford University
sradwan@stanford.edu

Michael C. Frank
Department of Psychology
Stanford University
mcfrank@stanford.edu

Daniel L. K. Yamins
Department of Psychology
Stanford University
yamins@stanford.edu

Judith E. Fan
Department of Psychology
UC San Diego
jefan@ucsd.edu

Abstract

Curiosity drives much of human behavior, but its open-ended nature makes it hard to study in the laboratory. Moreover, computational theories of curiosity – models of how intrinsic motivation promotes complex behaviors – have been challenging to test because of technical limits. To circumvent this problem, we develop a new way to assess intrinsic motivation for building: we assume people build what they find interesting, so we asked them to rate the "interestingness" of visual stimuli – in this case, simple block towers. Adults gave a range of ratings to towers built by children, with taller towers rated higher. To probe interestingness further, we developed controlled tower stimuli in a simulated 3D environment. While tower height predicted much of the variation in ratings, people also favored more precarious towers, as inferred from geometric features and simulated dynamics. These ratings and features therefore give a clear target for computational accounts of curiosity to explain.

Keywords: curiosity; play; intrinsic motivation; intuitive physics; visual abstraction

Introduction

Given a set of blocks, toddlers within the first 17-32 months of life readily stack them to produce block towers and other stable physical configurations (Bullock & Lütkenhaus, 1988). By 4-8 years of age children are capable of reasoning about how existing towers are built (Dietz, Landay, & Gweon, 2019; Dietz et al., 2019), an ability that continues to be refined into adulthood (McCarthy, Kirsh, & Fan, 2020). Building towers from blocks is perhaps the most basic version of our more general capacity to create new structures, from tall buildings to novel molecules and complex software programs.

While we often undertake a building project to achieve an instrumental goal (e.g. creating shelter or medicine), building things can also be *fun*. Following one's curiosity to imagine and construct alternative configurations of the world has long been recognized as a critical component of human learning (James, 1983) and cognitive development (Gopnik, Meltzoff, & Kuhl, 1999; Piaget & Cook, 1952). However, there are few theories that explain what intrinsically drives people to explore and play with their environment (Kidd & Hayden, 2015). A satisfying theory should account for why, in a simple setting like a room full of blocks, people build structures instead of flinging objects randomly, and why they choose to build some structures over others.

Computational models of behavior can be used to articulate quantitatively precise theories of intrinsic motivation. For example, when artificial agents are "motivated" to create scenarios whose dynamics the agents cannot easily predict, they both begin to perform nonrandom behaviors (e.g., preferentially attending to movable objects, object gathering, and smashing objects together) and to better recognize objects in their environment (Haber, Mrowca, Wang, Fei-Fei, & Yamins, 2018). Other forms of artificial curiosity (AC) have been proposed, in various domains, to account for the emergence of more complex behaviors and to drive learning about the world (Schmidhuber, 1991, 2010; Aubret, Matignon, & Hassas, 2019). These include recent methods for formalizing learning progress (Oudeyer & Kaplan, 2009; Oudeyer, Baranes, & Kaplan, 2013; Kim, Sano, De Freitas, Haber, & Yamins, 2020), for creating scenarios that violate expectations (Pathak, Agrawal, Efros, & Darrell, 2017), and for novelty-seeking (Burda, Edwards, Storkey, & Klimov, 2018).

To date, though, none of these types of AC has induced artificial agents to perform more elaborate object stacking or physical assembly behaviors in a physically realistic setting. One reason this could be is due to technical difficulties getting artificial agents to perform *any* complex behaviors using leading methods; state-of-the-art reinforcement learning agents require millions of trials just to learn short sequences of a few possible actions in simulated 2D environments (Burda, Edwards, Pathak, et al., 2018). Moreover, it has been technically infeasible to simulate interactions between many objects in a physically realistic 3D environment, although the ability to do so would enable much stronger comparisons between natural human behavior and that of artificial agents. An alternative reason that existing AC proposals may have failed so far is because they are wrong. For example, the drive to seek out physical scenarios that are *hard* to predict may not be sufficient to explain the emergence of complex physical assembly behavior, given that the dynamics of physical structures are in some ways *easy* to predict, especially if they remain static over time. Thus, there is a strong need to test theories of AC without the confounding influence of current technical shortcomings in artificial agent behavior.

The present project has two goals: The first is to develop an alternative approach for testing ideas about artificial cu-

* denotes equal contribution

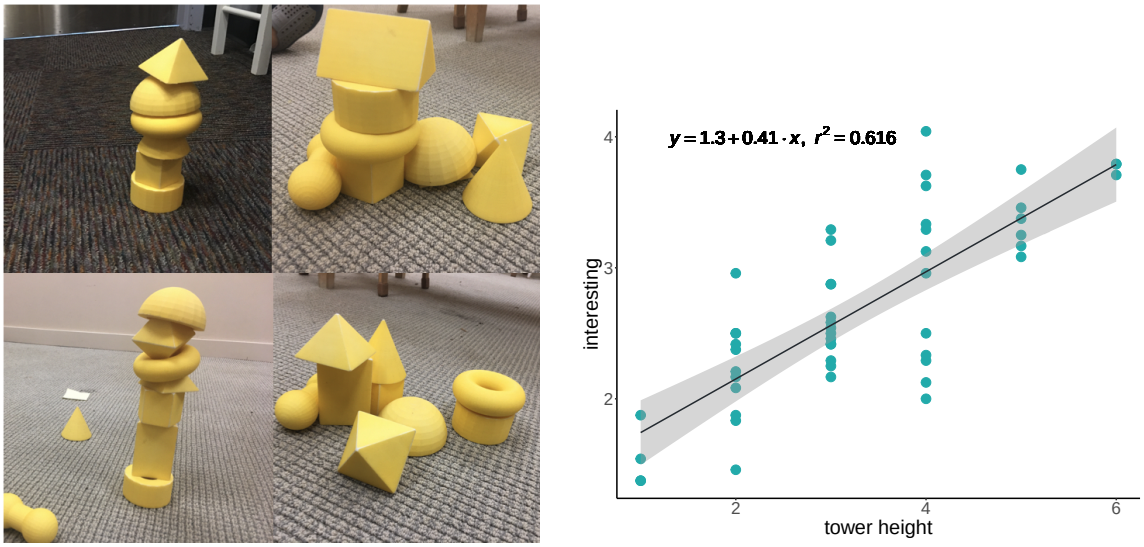

Figure 1: Left: Example structures built by children in Experiment 1. Each child was given 9 different blocks and one minute to construct a "cool tower". Right: Mean interesting ratings for each tower by tower height. Taller towers were consistently judged as more interesting by adult raters.

riosity against human behavior. Instead of trying to measure complex artificial construction behaviors and comparing them with what people do, we characterize *what people find interesting* about pre-built structures. If we assume people choose to imagine and build what they find interesting, then the same visual stimuli that are interesting to people should also be interesting to AC models – that is, they should elicit a strong feedback signal to construct these over "boring" stimuli. Thus, judgments of "interestingness" provide an observable that current and hypothetical forms of AC should quantitatively account for.

The second goal is to propose quantitative features of block towers that can predict interestingness judgments of these stimuli – and thereby indicate what makes a tower desirable to build. While such features may not immediately generalize to theories of AC in other physical domains, any satisfactory and general form of AC should at least explain why these particular features are interesting in the tower domain. Furthermore, simple heuristic models of tower interestingness may suggest what features a successful AC model should be sensitive to – for instance, the number of discrete objects arranged in a stable configuration.

We performed two experiments to probe judgments about interesting block towers. In the first, children were instructed to "build a cool tower" with a set of variably shaped blocks. This was conducted at the end of a larger experiment which explored how children selectively drop and collide these same objects. This exploratory study allowed us to observe the types of structures that children were inherently motivated to build and to measure which of these towers adults found interesting, testing our core assumption that interestingness judgments can provide us with clues about assembly behaviors. The results of this first study suggested that interestingness was related to tower height; however, they did not indicate whether tall towers were interesting *per se* or whether they were interesting because they contained a more diverse set of blocks, were more precarious (and therefore represented harder feats of construction), or simply more visually pleasing.

To distinguish these possibilities, our second experiment asked for both interestingness and stability ratings on a set of parametrically controlled tower images, which were generated in the physically realistic, Unity-based simulation environment ThreeDWorld (Gan et al., 2020). This study directly manipulated the height and stability of towers, and revealed that both of these factors led to more interesting structures. By contrast, differences in color and viewpoint did not impact ratings of interestingness, suggesting that these judgments were primarily about towers' physical properties, rather than incidental aspects of their visual appearance. Finally, we found that we could predict perceptual judgments with simple heuristic models of tower precariousness, computed from the ground truth states of the simulated towers and counterfactual physical dynamics (i.e., stochastically shifting blocks horizontally.) Thus, people may base their judgments of towers on particular inferences about their static geometry and possible dynamics, with the most interesting towers being taller and on the verge of falling over.

The fact that judgments of towers are both reliable and predictable suggests that current and future AC models should register more interesting towers as more worthy of building. Our results also hint that successful AC models will need to represent physical, not just visual, features of their environment, and general forms of intrinsic motivation should "reduce to" the particular physical feature combinations identified here when placed in a block tower-building environment.

Experiment 1: What kind of block towers are children motivated to build?

Our preliminary study explores the properties of towers children find intrinsically motivating to build, which could further lend insight to a common origin of physical "interestingness" judgements among adults. We collected images of towers built by children using plastic blocks in an open-ended, semi-controlled assembly task. Then we elicited adult ratings of how interesting these tower structures were.

Methods

Participants We recruited 53 children from the Children's Discovery Museum of San Jose and Bing Nursery School. Participant exclusions were made based on cases where i) child received help from researcher during tower assembly task or ii) the parent did not consent for video recording of study. After exclusions, results from 50 children were analyzed, including 6 2-year-olds, 17 3-year-olds, 15 4-year-olds, 10 5-year-olds, and 2 6-year-olds.

Stimuli Stimuli were 3D-printed plastic objects produced using Blender 3D-modeling software. The nine objects were: bowl, cone, dumbbell, octahedron, pentagonal prism, pipe, pyramid, torus, and triangular prism. The printed objects were all yellow, rigid plastic material and designed to fit comfortably in a child's hand (dimension range: 3.8-10.1 cm). Examples of these blocks can be seen in the sample towers shown on the left of Figure 1.

Procedure We asked the child to "make a cool tower" with any of the nine toy blocks for about one minute. A video camera was used to record the play session from an angle above the tower assembly space. Once the child completed the task, a researcher took a photo of the final tower to be saved for annotation.

Results

The average height of the constructed towers was 3.43 blocks, 95% CI [3.10, 3.76]. We examined tower height as a function of age, and found that older children tend to build taller towers ($r(44) = .32$, $p = .028$).

To investigate what made these towers "interesting" to adult viewers, we recruited 25 adults on Prolific to provide interestingness judgments ranging from 1 (not interesting at all) to 5 (extremely interesting). We estimated the effect tower height had on the rated interestingness using linear mixed effects models (LME) with a single fixed effect for tower height and random effects for each rater and tower. We found that the height of the tower was indeed a strong predictor of the rated interestingess; ($b = 0.377$, $t = 6.823$, $p < 0.001$). While it is possible that tall towers are interesting by virtue of being tall, we hypothesized that the height and stability interact to predict what adults find interesting to look at. To test this, we designed a tower rating experiment based on computer generated towers inspired by the child-built structures.

Experiment 2: What kind of block towers do adults find most interesting?

From the results of Experiment 1, we were motivated to more systematically investigate the relationship between tower height and stability in humans' perceptions of interestingness.

Methods

Participants We recruited 180 US adults via the online platform Prolific who were randomly assigned to either a stability or interestingess conditions. Prior to data collection, we determined to exclude any participants who did not complete the entire study, or who failed to pass two attention checks presented during the experiment. In total 17 were excluded, resulting in 74 participants in the interesting condition, and 93 in the stability condition.

Stimuli The towers were generated using the ThreeDWorld physics environment (Gan et al., 2020). Each tower was comprised of cube blocks that were stacked vertically, and systemmatically generated to vary along the horizontal and vertical axes. Variation along the horizontal axis was determined according to a jitter in the x–position of each block, and variation in the vertical direction was determined by the number of blocks in the tower. We used a 3x3 design with three levels of x–jitter ("low", "medium", and "high") and three possible numbers of blocks (2, 4, or 8 blocks). Jitter was defined by the variance of x–positions of each of the blocks. The x–coordinate of each block in the "low", "med", and "high" condition towers were sampled from a uniform distribution ranging from 0, 1/3, and 1/2 block widths from center, respectively. This sampling method yielded towers whose variance along the x–axis subtly increased across conditions. Within each condition we generated 8 towers from two different viewpoints yielding 144 target towers. The left side of Figure 2 shows example stimuli in each of the block number/jitter conditions, rendered from the upper right viewpoint.

Procedure Participants were randomly assigned to provide ratings on either tower interestingness or stability for 144 tower images. The order of the towers was randomized and each tower was shown individually. Participants rated the tower on a 1-5 scale ranging from "not interesting (stable) at all" to "extremely interesting (stable)."

Results

Our goal was to characterize which features of pre-built structures people find interesting. Because interestingness has not previously been studied in this domain, we first measured its reliability and compared it to that of stability ratings, which have been more widely studied. We then conducted a series of model comparisons to test how much these interestingness judgments rely on purely visual properties (i.e., color or viewing angle), geometric properties (i.e., height), or more complex physical properties (i.e., precariousness, as revealed by counterfactual simulations of physical dynamics).

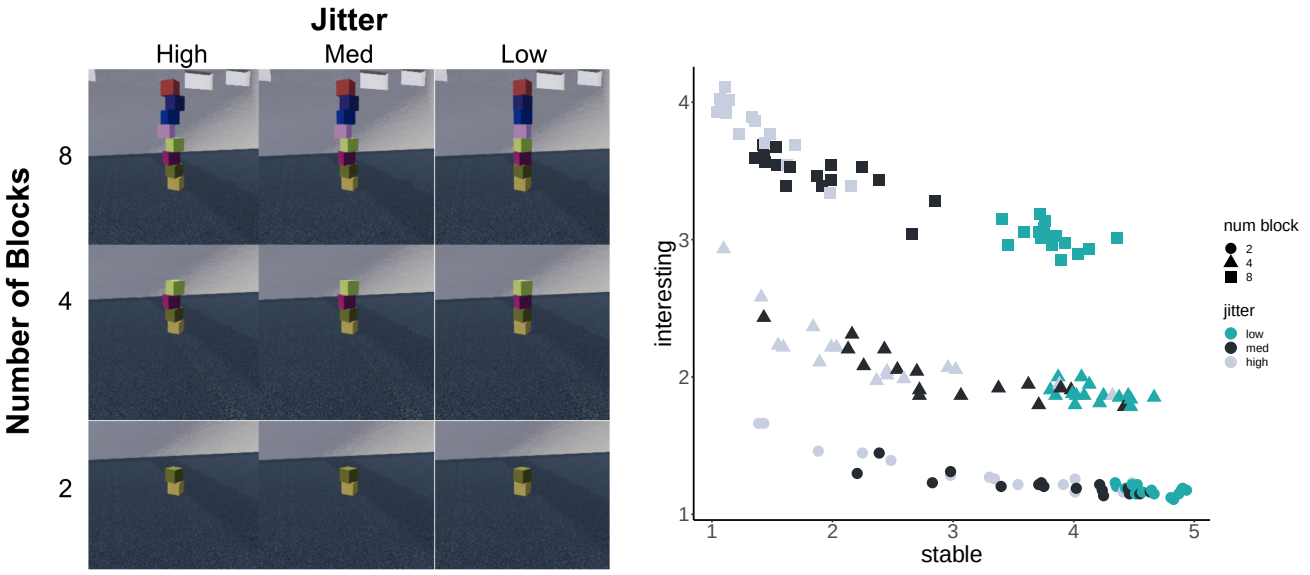

Figure 2: Left: Example stimuli generated in the ThreeDWorld physics environment. We utilized a 3x3 stimulus design where we systemmatically varied height (2,4,8 blocks) and x–jitter (low, med, and high jitter). Each tower was also rendered from two viewpoints, from the lower left and upper right. Right: Mean interesting and stability ratings for each tower. Towers with more blocks and greater jitter were rated as more interesting, and there was a significant interaction between the two variables. Viewpoint did not significantly predict either measure.

Interestingness judgments are as reliable as stability judgments. We first compared how reliable ratings across participants were for the stability *versus* interestingness judgments. Both conditions had very high reliability in average tower rating across participants; mean Spearman-Brown corrected correlation coefficient for split-halves of participants in the "stable" condition (0.997 ± 0.001) and the "interesting" condition (0.996 ± 0.001). We also calculated the proportion of responses for each tower that matched the mode response for that tower. Again, there was strong agreement across individuals in both conditions; proportion of modal agreement 51.7 ± 2.7 and 54.2 ± 3.1 for stable and interesting judgments, respectively. Finally, the average standard deviation of responses within each tower was similar across conditions; 0.79 and 0.77 for stable and interesting, respectively. These results suggest that interestingness ratings are highly reliable across participants and comparable in reliability to stability ratings.

Physical tower parameters account for interestingness judgments. Assessing stability calls for *physical* inferences about a structure and its components, rather than mere judgments of low-level visual properties like viewing angle, color, texture, and apparent (*versus* actual) size. Insofar as stability and interestingness are related, we hypothesized that a tower's generative physical parameters, (1) number of blocks/height ("height" in Figure 4) and (2) amount of jitter in the block positions ("jitter" in Figure 4), would predict interestingness ratings better than the purely visual parameter, (3) viewpoint ("viewpoint" in Figure 4). The first three rows of Figure 4 shows the R^2 values for LME models with random effects for participant and tower, and fixed effects for viewpoint ($R^2 = 0.000$), jitter ($R^2 = 0.017$), and height ($R^2 = 0.515$). Consistent with Experiment 1, height supplied the majority of predictive power in interesting ratings. To test whether (2) and (3) further improved performance above (1) alone, we conducted likelihood ratio tests on a sequence of LME models. We found that, as predicted, a model with an interaction between (1) and (2) performs significantly better than a simpler model with (1) and (2) as additive fixed effects ($\chi^2(2) = 32.699, p < 0.001$). The right side of Figure 2 shows the interaction of (1) and (2) in interestingness and stability ratings. Adding (3) as a fixed effect to the interaction model did not significantly improve model performance ($\chi^2(1) = 1.726, p = 0.189$), consistent with our hypothesis that both types of judgment would be insensitive to physically irrelevant properties of the tower stimuli.

Judgments are based on visual inference of physically relevant tower properties. The interaction of height and jitter strongly suggested that people make interestingness judgments by inferring physical features of the scene they are viewing and performing some (possibly complex) computation on those features. To test this idea, we created new models for predicting ratings from various components of the tower's visible silhouette and its "ground truth" physical state in the simulator (rather than from the discrete stimulus categories above, which participants did not know about.) Specifically, we calculated the height of each tower and the variance in the horizontal positions of its blocks ("variance" in Figure 4), properties that can be visually estimated to some degree; this approach circumvents the question of how accurately people can actually estimate the physical state of

a set of objects from visual input, so the predictive power of our models should be considered an upper bound. We again found that, even with image-computable features, an interaction between height and x-variance ($b = 1.906, t = 6.451, p < 0.001$) – accounted for a significant portion of variance in interestingness ratings. Likewise, adding a non-physical visual property, the mean RGB color intensity across blocks ("color" in Figure 4), did not improve this model ($\chi^2(1) = 1.122, p = 0.289$).

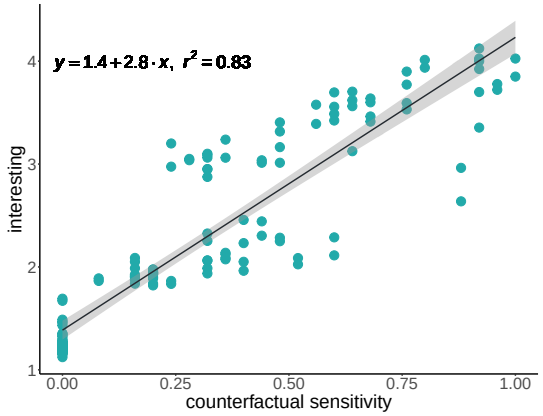

Figure 3: A counterfactual analysis of tower stability. Original towers were generatively resampled with noisy perturbations to block placement. The x-axis shows the "counterfactual sensitivity" – the percent of counterfactual towers that fell; and y-axis shows the mean interestingness rating for each tower.

A counterfactual model of surprisal predicts interestingness. The models explored so far appeal directly to properties specific to our tower stimuli, namely their height, their arrangement of blocks, and their colors. While these models explain a substantial proportion of the variance in interestingness judgments, they are heuristics that do not apply to more general physical stimuli and therefore could not act directly as intrinsic motivation signals. A more general model of interestingess would need to explain *why* taller and more precarious towers are more interesting without direct reference to their being towers or to tower-specific properties. Inspired by work on dynamical simulation as a model of judging physical stimuli (Battaglia, Hamrick, & Tenenbaum, 2013) and by the fact that towers judged less stable were also judged more interesting (Figure 2), we developed a "counterfactual sensitivity" model of interestingness ("counterfactual sensitivity" in Figure 4). For each of the original towers, we created 25 counterfactual versions in which each block had a 50% chance of being shifted from its original location in a random horizontal direction and with a random magnitude (sampled from a normal distribution with mean 0 and standard deviation equal to 1/4 of the block width.) The counterfactual sensitivity of each tower was then computed as the proportion of these alternative towers that, after forward simulation, reached a different static equilibrium from the original – that is, fell over.

Remarkably, this single feature explained most of the variance in mean tower interestingness ratings (Figure 3). Consistent with our hypothesis, as counterfactual sensitivity increased (high proportion of counterfactual towers fell over), mean interestingness ratings increased ($b = 2.846, t = 18.716, p < 0.001$). We also found that a LME model that adds this counterfactual measure outperformed the height/x-variance interaction only model ($\chi^2(1) = 13.298, p < 0.001$). Crucially, constructing this model does not depend on the stimuli being towers: *any* arrangement of physical objects could be counterfactually perturbed and simulated in this way, then assessed for whether the outcome was the same or different from what was observed. Thus, capturing a probabilistic notion of "how the scene *might* have been" (Battaglia et al., 2013) could provide a more general principle underlying interestingness judgments. This formulation also closely mirrors some accounts of artificial curiosity, in which agents find it intrinsically rewarding to see outcomes that violate their expectations of how a scene will unfold (Achiam & Sastry, 2017; Haber et al., 2018). In the present domain, this measure of surprisal captures a notion of tower precariousness, but it could be extended to explain the interestingness of other entity types (e.g. nonrigid bodies, fluids) and physical scenarios (e.g. collisions, drops). In future work we will explore the extent to which dynamical simulation and inference can capture what makes other domains interesting.

Discussion

Studying intrinsic motivation in the laboratory or in simulated environments has been challenging because curiosity most naturally arises in complex, open-ended contexts where current computational models struggle. In this work we began to address this issue by proposing "interestingness" as a measure of intrinsic motivation that can be assessed through perceptual judgments.

The particular stimulus features that people found interesting here suggest clear ways to extend the set of judgments. Tower height was the dominant predictor of tower interestingness; properties related to "precariousness" – but not physically irrelevant visual properties – played a second-order role. These simple features however do not explain all explainable variance, and indeed even including the mean stability ratings for each tower left much variance to be explained. We also showed that models inspired by counterfactual simulation can capture much of the variance in mean interestingness ratings, suggesting there may be domain general relationships between expectation violation (what *might* have been) and what people find interesting.

Further work will be required to find out what else determines interestingess. In our experiments, tower height was confounded with the number of blocks (since these were of uniform size), so future work should test whether equally tall towers made of variable numbers of blocks are more or less

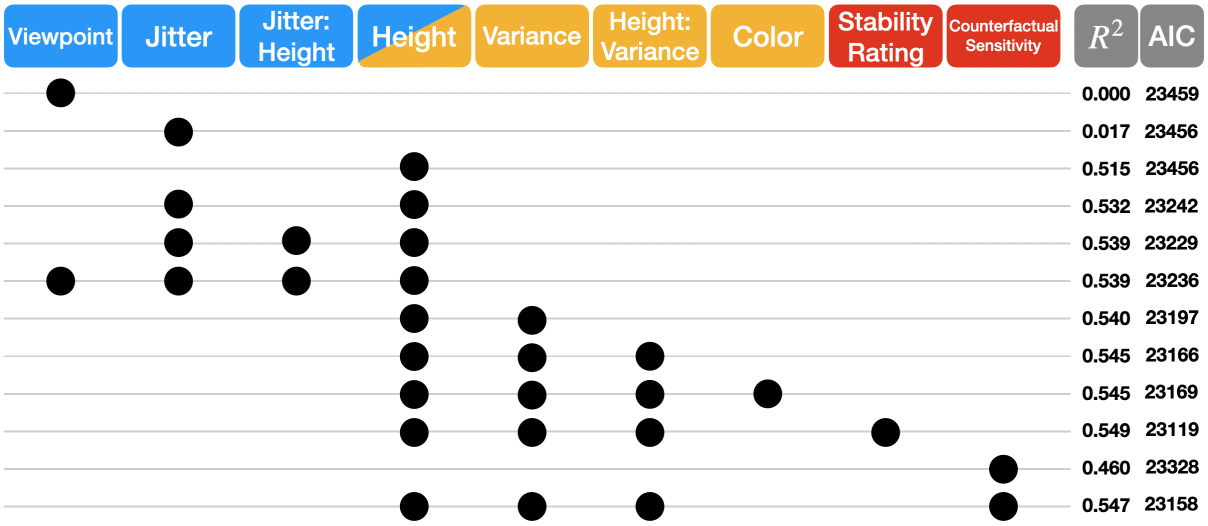

Figure 4: Linear mixed effects model comparisons; columns denote fixed effects and rows with a circle denote that effect was included in the model; all models included random effects for participant and tower. Interactions between fixed effects are denoted with a ":". Blue columns are experimental conditions; yellow columns are visually computable features (tower height is both an experimental and image computable feature); red columns are features that require physical inference. Marginal R^2 and AIC are computed for each model.

interesting. The "cool towers" built by children in Experiment 1 also hint that including non-cubic blocks may produce even more interesting towers, as they allow for both new precarious shape combinations and non-rectilinear geometries. By expanding the set of stimulus variables and range of perceptual judgments, these further experiments will give an increasingly precise target for artificially curious models to hit.

Our results indicate that interestingness judgments could be used to compare people and computational accounts of curiosity. Because judgments spanned a range of interestingness values and people generally agreed about which tower stimuli were interesting, these data are rich targets for such accounts to explain: if a given model of artificial curiosity observed or imagined a highly rated tower and found it boring (i.e., the tower did not elicit a strong construction-motivating signal), then the model would not be a satisfactory explanation of human curiosity and exploratory behavior. Testing existing forms of artificial curiosity on the data collected here is therefore a critical next step in this line of work.

Already, though, our results raise several possible reasons that current artificial agents do not build elaborate structures. Taller towers are dramatically more interesting than two-block "towers" here, but artificial agents struggle to learn even to stack one object on top of another when placed in a physically realistic environment and given a realistically large set of plans to choose from (Haber et al., 2018). Thus, whether or not they would perceptually judge a tall tower as worth building, the current generation of reinforcement learning algorithms is likely hampered more directly by technical failure to get off the ground (Curtis, Xin, Arumugam, Feigelis, & Yamins, 2020). Models most interested in situations that violate their model of the world (Schmidhuber, 1991; Pathak et al., 2017; Haber et al., 2018) may never *encounter* tall towers, let alone acquire expectations about them; models of artificial curiosity focused on novelty (Burda, Edwards, Storkey, & Klimov, 2018) likewise will not come across tall towers by chance. Intuitively, methods that involve setting "interesting" goals for oneself (Campero et al., 2020) and tracking learning progress (Kim et al., 2020) might nudge agents toward building tall towers, but to be useful models they will have to explain *why* these structures are interesting – in other words, why general curiosity "reduces to" building tall towers in this simple environment. All of these considerations further stress the need to test theories of artificial curiosity as independently as possible from models of motor behavior and planning.

Finally, our findings suggest a few ingredients that may be important for an artificially curious agent. People appear to ignore physically irrelevant properties of stimuli (viewpoint, color) and apply physical intuition about stability in judging towers. As such, computational models that abstract visual inputs into physical objects (Bear et al., 2020) and simulate their dynamical behavior (Battaglia et al., 2013; Li et al., 2020) may be necessary, though not sufficient (Curtis et al., 2020), for getting artificial agents to make human-like judgments and building decisions. Directly comparing models with these ingredients to human judgments will test whether they give a quantitatively better account of "interestingness" than simpler models of visual processing. In developing a new approach to measuring intrinsic motivation, our broader aim is to better understand what common principles underlie the rich and complex behaviors that both adults and children exhibit in realistic physical environments.

Acknowledgements

C.H. is supported by a DoD NDSEG Fellowship. D.M.B. is supported by the Wu Tsai Neurosciences Institute and a Biogen Fellowship from the Life Sciences Research Foundation. This work was also supported by NSF CAREER Award #2047191 to J.E.F.

All code and materials available at:
`https://github.com/cogtoolslab/curiotower`

References

Achiam, J., & Sastry, S. (2017). Surprise-based intrinsic motivation for deep reinforcement learning. *arXiv preprint arXiv:1703.01732*.

Aubret, A., Matignon, L., & Hassas, S. (2019). *A survey on intrinsic motivation in reinforcement learning.*

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.

Bear, D. M., Fan, C., Mrowca, D., Li, Y., Alter, S., Nayebi, A., ... others (2020). Learning physical graph representations from visual scenes. *arXiv preprint arXiv:2006.12373*.

Bullock, M., & Lütkenhaus, P. (1988). The development of volitional behavior in the toddler years. *Child Development*, 664–674.

Burda, Y., Edwards, H., Pathak, D., Storkey, A., Darrell, T., & Efros, A. A. (2018). Large-scale study of curiosity-driven learning. *arXiv preprint arXiv:1808.04355*.

Burda, Y., Edwards, H., Storkey, A., & Klimov, O. (2018). Exploration by random network distillation. *arXiv preprint arXiv:1810.12894*.

Campero, A., Raileanu, R., Küttler, H., Tenenbaum, J. B., Rocktäschel, T., & Grefenstette, E. (2020). *Learning with amigo: Adversarially motivated intrinsic goals.*

Curtis, A., Xin, M., Arumugam, D., Feigelis, K., & Yamins, D. (2020). *Flexible and efficient long-range planning through curious exploration.*

Dietz, G., Landay, J. A., & Gweon, H. (2019). Building blocks of computational thinking: Young children's developing capacities for problem decomposition. In *Cogsci* (pp. 1647–1653).

Gan, C., Schwartz, J., Alter, S., Schrimpf, M., Traer, J., De Freitas, J., ... others (2020). Threedworld: A platform for interactive multi-modal physical simulation. *arXiv preprint arXiv:2007.04954*.

Gopnik, A., Meltzoff, A. N., & Kuhl, P. K. (1999). *The scientist in the crib: Minds, brains, and how children learn.* William Morrow & Co.

Haber, N., Mrowca, D., Wang, S., Fei-Fei, L. F., & Yamins, D. L. (2018). Learning to play with intrinsically-motivated, self-aware agents. In *Advances in neural information processing systems* (pp. 8388–8399).

James, W. (1983). *Talks to teachers on psychology and to students on some of life's ideals* (Vol. 12). Harvard University Press.

Kidd, C., & Hayden, B. Y. (2015). The psychology and neuroscience of curiosity. *Neuron*, *88*(3), 449–460.

Kim, K., Sano, M., De Freitas, J., Haber, N., & Yamins, D. (2020). Active world model learning with progress curiosity. In *International conference on machine learning* (pp. 5306–5315).

Li, Y., Lin, T., Yi, K., Bear, D., Yamins, D., Wu, J., ... Torralba, A. (2020). Visual grounding of learned physical models. In *International conference on machine learning* (pp. 5927–5936).

McCarthy, W., Kirsh, D., & Fan, J. (2020). Learning to build physical structures better over time. In *Cogsci.*

Oudeyer, P.-Y., Baranes, A., & Kaplan, F. (2013). Intrinsically motivated learning of real-world sensorimotor skills with developmental constraints. In *Intrinsically motivated learning in natural and artificial systems* (pp. 303–365). Springer.

Oudeyer, P.-Y., & Kaplan, F. (2009). What is intrinsic motivation? a typology of computational approaches. *Frontiers in neurorobotics*, *1*, 6.

Pathak, D., Agrawal, P., Efros, A. A., & Darrell, T. (2017). Curiosity-driven exploration by self-supervised prediction. In *International conference on machine learning* (pp. 2778–2787).

Piaget, J., & Cook, M. (1952). *The origins of intelligence in children* (Vol. 8) (No. 5). International Universities Press New York.

Schmidhuber, J. (1991). Curious model-building control systems. In *Proc. international joint conference on neural networks* (pp. 1458–1463).

Schmidhuber, J. (2010). Formal theory of creativity, fun, and intrinsic motivation (1990–2010). *IEEE Transactions on Autonomous Mental Development*, *2*(3), 230–247.