Faulty Towers: A counterfactual simulation model of physical support

Tobias Gerstenberg, Liang Zhou, Kevin A. Smith & Joshua B. Tenenbaum
{tger, zhoul, k2smith, jbt}@mit.edu
Brain and Cognitive Sciences, Massachusetts Institute of Technology

Abstract

an abstract ... **Keywords:**

Introduction

When we look at a physical scene, such as the towers shown in Figure 1, we don't just see a pile of bricks. We also have a sense for how stable the towers are, what would happen if the table got bumped in one direction or another, and what the relative masses of different bricks must be given that the tower is stable (Battaglia, Hamrick, & Tenenbaum, 2013; Hamrick, Battaglia, Griffiths, & Tenenbaum, 2016). In this paper, we show that people can not only gage the stability of the tower overall, but also judge the extent to which different bricks individually contribute to the tower's stability. We develop a *counterfactual simulation model* (CSM) of physical support which determines a brick's causal responsibility for the tower's stability by simulating what would happen if the brick was removed.

TG: maybe talk about recent successes in deep learning with these tasks?; peter's new paper?!

In previous work, we showed how the counterfactual simulation model explains people's causal judgments about dynamic collision events (Gerstenberg, Goodman, Lagnado, & Tenenbaum, 2012, 2014, 2015; Gerstenberg & Tenenbaum, 2016). In these experiments, participants saw collisions between billiard balls and they were asked to evaluate to what extent one ball had caused another ball to go through a gate in a wall (or prevented the ball from going through). The CSM assumes that people reach this judgment by comparing what actually happened with what would have happened in a counterfactual situation in which the candidate cause had been removed from the scene (or perturbed). In line with the CSM, the results of the experiments showed that there was a very close correspondence between the counterfactual judgments of one group of participants, and the causal judgments of another group. As predicted by the model, participants' cause and prevention judgments increased the more certain they were that the outcome would have been different if the candidate cause had been removed from the scene. The CSM not only predicts participants' causal judgments to a high degree of quantitative accuracy, it also captures the cognitive processes by which participants reach their judgments. Participants' eye-movements reveal how they spontaneously anticipate what would have happened in the relevant counterfactual situation (Gerstenberg, Peterson, Goodman, Lagnado, & Tenenbaum, resubmitted). The CSM makes the strong prediction that counterfactual simulation forms a necessary part of how people make causal judgments, and that no adequate account of people's causal judgments about particular events can be developed that does not rely on counterfactuals (cf. Wolff, 2007). Thus far, however, the CSM has only been applied to modeling causal judgments about dynamic collision events. Here, we demonstrate the generality of the account by showing how the model naturally handles judgments about physical support.

Judging support is different from judging causation in several ways. For example, most philosophical approaches to causation take the causal relata (i.e. the things that do the causing) to be events (Halpern, 2016; Paul & Hall, 2013). It is the player's kicked that caused the ball to go in the goal. However, when we consider the extent to which a particular brick is causally responsible for the tower's stability, nothing actually happens. The tower is just sitting there – there are no events. So, rather than defining a counterfactual operation on events, CSM considers what would happen to the tower if the brick was removed. The more certain we are that the tower would collapse, the more responsible the brick for its stability.

The road map for the rest of the papers is as follows: We will first discuss the relevant background literature. Then, we present the CSM in detail. We will test the model with two experiments that ask some participants to make hypothetical judgments, and others to evaluate causal responsibility. We end by discussing limitations of the current approach, and directions for future research.

TG: signpost: what is the right counterfactual model? reference to Lewis (counterfactual most similar to the actual world); no mismatch in emphasis between introduction and rest of the paper

A counterfactual simulation model of physical support

In our experiments, we ask participants how responsible the black brick is for the red bricks staying on the table. To derive predictions from the CSM we need to determine (1) what counterfactual situation to consider, and (2) how to simulate what would happen in that situation. We assume that when judging responsibility, participants consider a counterfactual situation in which the black brick was removed. Participants then use their intuitive understanding of physics to mentally simulate what would happen in that situation.[1]

TG: footnote here needs to be updated

[1] We use the term *counterfactual* here broadly to refer to a possible world that
simulation model to highlight similarities to the model we developed for dynamic collision events. However, when judging
responsibility for support participants need not go back in time to think about what would have happened, but merely need to consider the *hypothetical* of what would happen if the brick was removed.

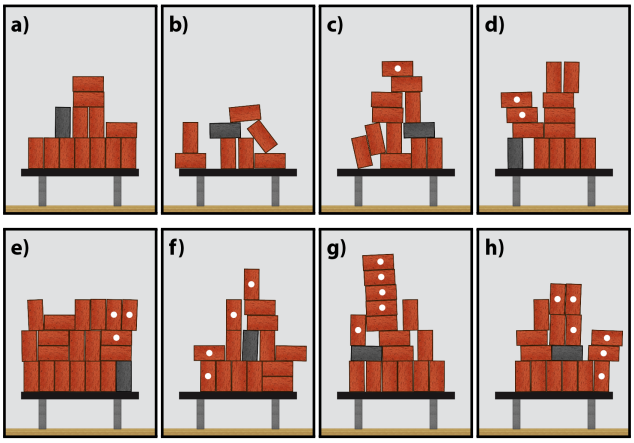


Figure 1: **Experiment 1**. Example stimuli. *Note*: Red bricks that would fall off the table if the black brick were removed (according to ground truth) are marked with a white dot at their center. The dots were not displayed in the actual experiment.

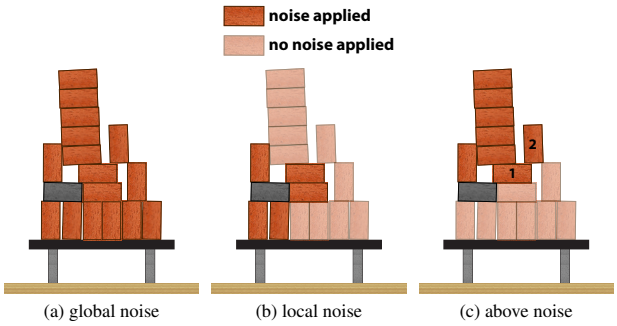


Figure 2: Schematic illustration of how different versions of the counterfactual simulation model apply noise when considering what would happen if the black brick were removed. The opacity of the red bricks indicates for each version of the model which bricks were subjected to noise.

TG: maybe add some more numbers to the loca-above model to explain why some bricks are not included

While some work suggests that people's understanding of the physical world is systematically biased (McCloskey, 1983), recent work has argued that some aspects of people's intuitive understanding of physics is well-described by making the assumption that people have an approximate simulation engine in their mind that is akin to a physics engine that generates physically realistic scenes (Battaglia et al., 2013; Lake, Ullman, Tenenbaum, & Gershman, 2016). Part of what makes these simulation engines in the mind "approximate" is that they assume that people's representation of a physical situation is uncertain. This uncertainty can come in many forms, such as perceptual uncertainty about the exact location of objects (Battaglia et al., 2013), dynamic uncertainty about how exactly an object will move (Smith & Vul, 2013), and uncertainty about latent physical parameters such as friction and mass (Sanborn, Mansinghka, & Griffiths, 2013). For example, in Battaglia et al.'s (2013) experiments, participants were asked to say whether a tower was stable, or in which direction a tower would fall. They modeled people's uncertainty by running many noisy simulations of the actual scene in which the position of the bricks was slightly perturbed. Participants' stability judgments were highly correlated with the average proportion of bricks that fell across these noisy simulations.

We contrast three different implementations of the CSM which differ in terms of how they capture people's uncertainty about what would have happened if the black brick had been removed. All models apply noise in the same way: as a small impulse to the red bricks very shortly after the black brick was removed. However, the models differ in which bricks they apply noise to when simulating the relevant counterfactual. Figure 2 illustrates how the three different models work. The *global noise* model applies a small impulse to all the bricks. The *local noise* model applies the impulse only to the red bricks that are directly in contact with the black brick. The *above noise* model applies noise only to bricks that are above the black brick and connected with it. To determine which bricks are connected it first looks at which bricks are directly in contact with the black brick *and* above it. It then recursively applies the same criterion to the identified red bricks. That is, it then looks at what other bricks are in contact with the identified red bricks, and above them. For example, brick 2 in Figure 2c is subject to noise since brick 1 is in contact and above the black brick, and brick 2 is in contact and above brick 1. The local noise models incorporate the assumption that only some aspects of the physical scene would be directly affected by the counterfactual intervention. Those aspects of a situation that are completely disconnected from the counterfactual intervention should remain exactly as they were. We test this assumption directly in Experiment 2 which features disjoint configurations of bricks (cf. Figure).

TG: say more about how this impulse is applied

TG: enter some more explanation by what we mean by "above"

TG: insert figure reference

Experiment 1

In the experiment, participants saw towers of bricks like the ones shown in Figure 1. Depending on the experimental condition, participants were asked to consider what would happen if the black brick wasn't there, or evaluate the extent to which the black brick is responsible that the red bricks stay on the table. In line with the CSM, we predicted that there would be a close relationship between counterfactual and responsibility judgments.

Methods

Design & Procedure The experiment had three experimental conditions that only differed in terms of the dependent measure. In the *prediction condition*, participants were asked to answer the question: "How many of the red bricks would fall off the table, if the black brick wasn't there?" Participants provided their answer on a sliding scale ranging from 0 to the number of red bricks present in the scene in steps of 1. In the *selection condition*, participants were asked to "Please click on the red bricks that would fall off either side of the table

if the black brick wasn't there." In the *responsibility condition*, participants were asked to answer the question: "How responsible is the black brick for the red bricks staying on the table?" Responses were provided on a sliding scale ranging from "not at all" (0) to "very much" (100).

The procedure for all three conditions was identical. Participants first received instructions about the task. They then saw a number of animations that showed 20 bricks being dropped on the table.[2] These animations were shown so as to familiarize participants with the relevant properties of the physical scene such as the bricks' mass, the friction between the bricks, as well as the table friction. Participants were allowed to proceed to the next stage once they had watched at least five animations.

After the familiarization, participants saw 42 images of different towers of bricks in randomized order (see Figure 1 for examples). The stimuli varied the number of bricks on the table (range = 7 to 20, M = 13.7, SD = 3.3), as well as the number of red bricks that would fall off the table if the black brick was removed (range = 0 to 6, M = 2, SD = 1.9). Participants' task differed depending on the condition as described above. Finally, participants were asked to give open-ended feedback about the task, and provided demographic information.

On average, the experiment took 9.86 (SD = 6.49), 15.71 (SD = 6.49), and 8.88 minutes (SD = 8.90) in the prediction, selection, and responsibility condition, respectively.

TG: if there is space/time, make a diagram that illustrates the three different conditions

Participants 121 participants (M_{age} = 34, SD_{age} = 12, 47 female) were recruited via Amazon Mechanical Turk with $N = 42$ in the prediction condition, $N = 38$ in the selection condition, and $N = 41$ in the responsibility condition. We excluded participants from further analysis based on their responses to the catch trial shown in Figure 1a. 11 participants in the prediction condition were excluded because they predicted that at least one red brick would fall. 6 participants in the responsibility condition were excluded because they gave a responsibility rating greater than 15 (on a scale from 0 to 100). No participants were excluded from the selection condition because no participant selected any of the bricks on the catch trial.

TG: rating greater than 15 sounds a little arbitrary...

Results

We will discuss the results from the *selection*, *prediction*, and *responsibility* conditions in turn.

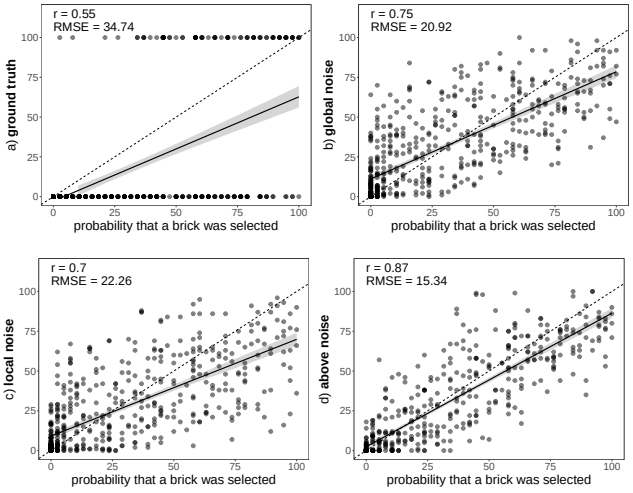


Figure 3: **Experiment 1**. Scatter plots showing the relationship between the empirical probability with which each brick was selected and the (a) ground truth as well as the predictions of the best-fitting (b) global noise model, (c) local noise model, and (d) above noise model.

Selection condition

TG: also test a regression model with features ...

We tested how well the three noise models described above captured participants' selections of which bricks would fall off the table if the black brick wasn't there (see Figure 2). For each model, we used maximum likelihood fitting to find the the noise parameter which predicts participants' selections best. For each setting of the noise parameter, we ran 100 simulations per stimulus and used the proportion of samples that each brick fell off the table in the noisy simulations to predict the probability that a given brick will be selected to fall by participants.[3]

Figure 8 shows how well the different noise models account for participants' selections. Overall, the *above noise* model accounts best for the data (cf. Table 1).

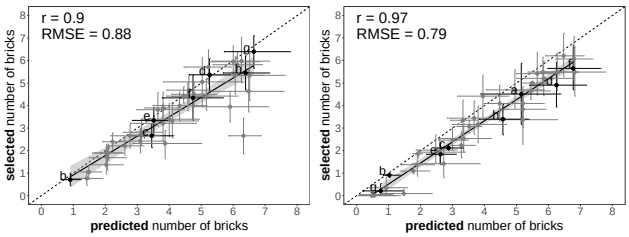
Table 1: Summary of model results for Experiments 1 and 2 as applied to the data in the *selection condition*.

	Experiment 1				Experiment 2			
model	r	RMSE	L	σ	r	RMSE	L	σ
truth	0.55	34.74	-21374	0	0.64	31.65	-22279	0
global	0.75	20.92	-9274	6.9	0.61	29.03	-14034	2.5
local	0.70	22.26	-9727	11.2	0.66	25.35	-12617	7.2
above	0.87	15.34	-8435	14.3	0.73	22.08	-11824	12.5

Note: r = Pearson correlation, RMSE = root mean squared error, L = log-likelihood of the data, σ = SD of the Gaussian from which the noise impulse is drawn.

[2] We used the same methodology to generate the stimuli for Experiment 1.

[3] Figure 8 gives an example for what these predictions look like for stimuli used in Experiment 2.



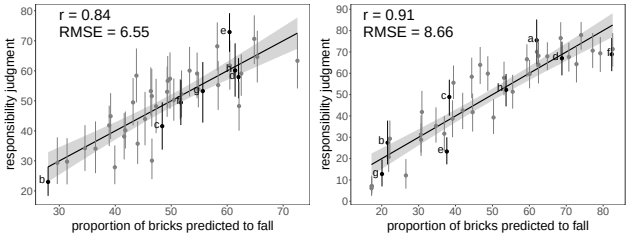
(a) **Experiment 1** (b) **Experiment 2**

Figure 4: Relationship between the predicted number of red bricks that would fall if the black brick wasn't there (prediction condition) and number of selected bricks that would fall (selection condition). *Note*: The letters refer to the examples shown in Figure 1 for Experiment 1, and Figure 6 for Experiment 2. Error bars denote bootstrapped 95% confidence intervals.

Prediction condition Figure 4 shows the relationship between the number of bricks predicted to fall, and the average number of bricks that participants selected in the selection condition. Overall, the two ways of probing participants' counterfactual simulations lead to very similar results. However, participants in the prediction condition predicted that more bricks would fall than participants in the selection condition selected (most of the data points are below the diagonal).

The noise model which best accounted for participants' selections, also accurately predicts participants' average judgments about how many bricks would fall with $r =$, RMSE $=$.

TG: needs to be updated once final simulations have been run



(a) **Experiment 1** (b) **Experiment 2**

Figure 5: Relationship between the predicted proportion of bricks that would fall if the black brick wasn't there and responsibility judgments. *Note*: The letters refer to the examples shown in Figure 1 for Experiment 1, and Figure 6 for Experiment 2. Error bars denote bootstrapped 95% confidence intervals.

TG: maybe have the same axis ranges?

Responsibility condition Figure 5a shows the relationship between the proportion of bricks that participants in the *prediction condition* believed would fall off the table if the black brick wasn't present in the scene, and participants' responsibility judgments. As predicted by the CSM, there was a very high quantitative relationship between counterfactual and responsibility judgments. This strongly suggests that participants evaluated a brick's responsibility by considering what proportion of bricks would have fallen off the table if the brick hadn't been there. When we use the proportion of bricks selected in the *selection condition* to predict participants responsibility judgments, we get a similarly good fit with $r = .78$, RMSE $= 7.65$. Finally, we can also use the best-fitting noise model to account for participants' responsibility judgments with $r = .78$, RMSE $= 7.65$. We will discuss the fact that responsibility judgments are more closely related to the prediction judgments compared to the selections in the General Discussion.

Discussion

The results of Experiment 1 support the predictions of the CSM. Most importantly, there was a very close relationship between the responsibility judgments of one group of participants, and the number of bricks that another group of participants predicted would fall if the black brick wasn't there. We contrasted three instantiations of the CSM the differ in the way in which they capture people's uncertainty about what would happen if the brick was removed. The results show that the *above model* explains participants' selections best.

Experiment 2

One limitation of Experiment 1 is that the predictions of the different noise models are highly correlated with each other. For example, the correlation between the *above* and *global* model is $r =$. Experiment 1 elicited participants' judgments for a wide array of different situations. In Experiment 2, we generated a smaller subset of situations that were more tightly controlled.

TG: maybe Liang can have another look through the participants' comments and mark interesting ones

TG: enter correlation here

- experiment that distinguishes between the different models
- illustrate the locality point by talking through 3 concrete examples

Methods

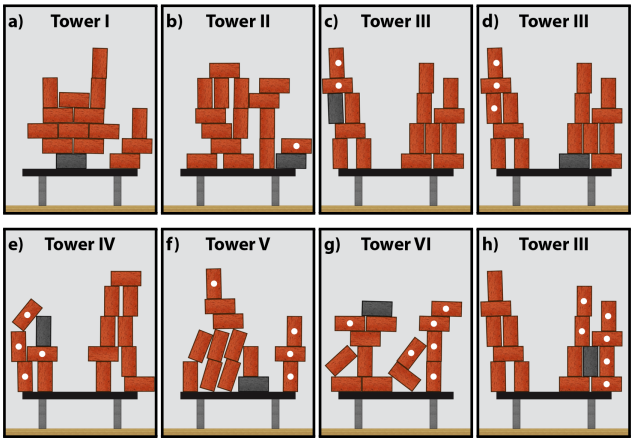


Figure 6: **Experiment 2**. Example stimuli. *Note*: The white dots indicate which bricks would fall if the black brick wasn't there. There were 6 different configurations of towers, and 7 different positions for the black brick in each tower, see c), d), and h).

TG: can remove the subtitles by adding text to the figures directly

Design & Procedure Figure 6 shows a selection of the stimuli. We generated six different tower configurations. For each tower, we chose seven different positions for the black brick. Figure 6c, d, and g show three examples for the position of the black brick for this configuration. The towers and brick positions were chosen such that removing the black brick would lead to 0–6 bricks falling off the table for the seven different instances of each tower configuration. Furthermore, as mentioned above, we included configurations which featured disjointed sets of bricks, such as Tower III and Tower IV.

The procedure was identical to that of Experiment 1. On average, the experiment took 11.57 (SD = 5.24), 13.04 (SD = 6.87), and 7.86 minutes (SD = 3.48) in the prediction, selection, and responsibility condition, respectively.

TG: maybe make the point that participants in the responsibility condition were pretty fast, so it's likely that they didn't think through each individual brick

Participants 129 participants (M_{age} = 36, SD_{age} = 11.3, 59 female) were recruited via Amazon Mechanical Turk with $N = 42$ in the prediction condition, $N = 44$ in the selection condition, and $N = 43$ in the responsibility condition. Using the same stimulus and exclusion criteria as in Experiment 1, 3 participants were removed in the prediction condition, 1 in the selection condition, and 3 in the responsibility condition.

Results

Again, we will discuss the results in the *selection*, *prediction*, and *responsibility* condition in turn.

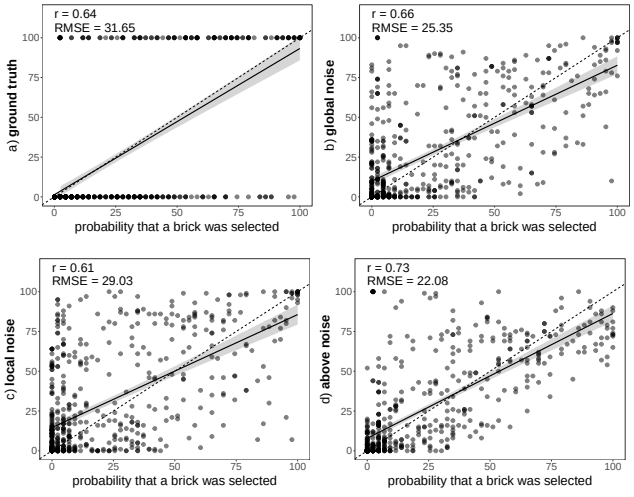


Figure 7: **Experiment 2**: Scatter plots showing the relationship between the empirical probability with which each brick was selected and the (a) ground truth as well as the predictions of the best-fitting (b) global noise model, (c) local noise model, and (d) above noise model.

Selection condition Figure 7 shows the correspondence between participants' brick selections and the predictions according to the ground truth as well as our three noise models as illustrated in Figure 2. Across all the stimuli, there were 564 bricks in total. Overall, the *above noise* model accounts for participants' selections best. Table 1.

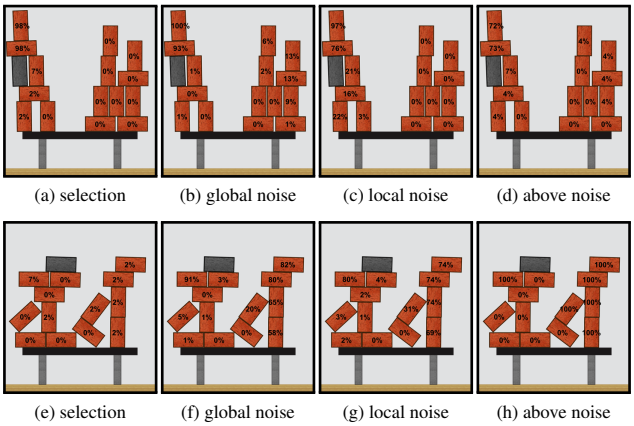


Figure 8: Empirical selection percentages for two different stimuli together with the predicted selection probabilities according to the different noise models.

Let us discuss the two situations shown in Figure 8 in some more detail. For the example shown in the top row, participants are confident that only the two bricks above the black one would fall (only very few participants selected any of the other bricks). As Figure 6c shows, participants' selections correspond closely to the ground truth in this case. Since the *global noise* model applies an impulse to all the bricks, it incorrectly predicts that bricks on the right would fall. In contrast, the *local noise* model correctly predicts that none of the bricks on the right will fall. However, since the model applies an impulse to all the bricks that are in contact with the black brick, it overpredicts that the bricks underneath the black brick would fall. The *above noise* model predicts participants' selection best in this case. It only assigns a small probability that any of the bricks on the right would fall (because sometimes the bricks on top of the black brick will fall towards the right), and a small probability that any of the bricks underneath the black brick would fall.

The example in the bottom row shows a situation where participants' selections didn't correspond to the ground truth. Here, the majority of participants believed that none of the bricks would fall if the black brick wasn't there. However, as Figure 6g shows, there are in fact six bricks that would fall according to the ground truth. When the black brick is removed, the two bricks directly underneath it fall to the left and right, and the one falling to the right pushes the stack of bricks on the right off the table. None of our noise models is able to capture participants' selections in this case. The *above noise* model does a particularly poor job for the simple reason that it doesn't apply any noise in this case. Since the black brick is on top, its predictions correspond to the ground truth. What this clearly shows is that our noise models don't yet completely capture participants' counterfactual simulations. We will discuss some ideas about how the improve the models in the General Discussion below.

Prediction condition

Responsibility condition

Discussion

- maybe show some examples for where the different noise models still get things wrong (and show images that display the percentage with which each brick was selected and predicted)

General Discussion

Most importantly, the results of both experiments showed that there was very close relationship between the counterfactual judgments of one group of participants, and the responsibility judgments of another group of participants.

- maybe combine small local noise with above noise ...
- make the point here that the model is "counterfactual" only in the lose sense; it doesn't require going back in time (so it's rather a "hypothetical" simulation model)
- maybe include a participant's comment here
  - "Playing angry birds helps on this, I can kind of visualize the falling bricks."
  - "How many bricks were on top of the black brick and what would happen if it were removed. If most of the bricks would fall off the table I figured it was more important. Also if the black brick was at the top I figured it was less important and had less impact."

Acknowledgments This work was supported by the Center for Brains, Minds & Machines (CBMM), funded by NSF STC award CCF-1231216 and by an ONR grant N00014-13-1-0333.

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.

Freyd, J. J. (1987). Dynamic mental representations. *Psychological Review*, *94*(4), 427–438.

Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2012). Noisy Newtons: Unifying process and dependency accounts of causal attribution. In N. Miyake, D. Peebles, & R. P. Cooper (Eds.), *Proceedings of the 34th Annual Conference of the Cognitive Science Society* (pp. 378–383). Austin, TX: Cognitive Science Society.

Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2014). From counterfactual simulation to causal judgment. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th Annual Conference of the Cognitive Science Society* (pp. 523–528). Austin, TX: Cognitive Science Society.

Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2015). How, whether, why: Causal judgments as counterfactual contrasts. In D. C. Noelle et al. (Eds.), *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (pp. 782–787). Austin, TX: Cognitive Science Society.

Gerstenberg, T., Peterson, M., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (resubmitted). Eye-tracking causality.

Gerstenberg, T., & Tenenbaum, J. B. (2016). Understanding "almost": Empirical and computational studies of near misses. In A. Papafragou, D. Grodner, D. Mirman, & J. C. Trueswell (Eds.), *Proceedings of the 38th Annual Conference of the Cognitive Science Society* (pp. 2777–2782). Austin, TX: Cognitive Science Society.

Halpern, J. Y. (2016). *Actual causality*. MIT Press.

Hamrick, J. B., Battaglia, P. W., Griffiths, T. L., & Tenenbaum, J. B. (2016). Inferring mass in complex scenes by mental simulation. *Cognition*, *157*, 61–76.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2016). Building machines that learn and think like people. *arXiv preprint arXiv:1604.00289*.

McCloskey, M. (1983). Naive theories of motion. In D. Gentner & A. L. Stevens (Eds.), *Mental models* (pp. 299–324). Erlbaum.

Paul, L. A., & Hall, N. (2013). *Causation: A user's guide*. Oxford University Press.

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., & Vul, E. (2013). Sources of uncertainty in intuitive physics. *Topics in Cognitive Science*, *5*(1), 185–199.

Wolff, P. (2007). Representing causation. *Journal of Experimental Psychology: General*, *136*(1), 82–111.

To do list

- "were removed" or "was removed"?
- run regression models with features
- look for best-fitting models for Experiment 2
- test a model that combines local with local-above noise

Literature review

Freyd (1987)

- sensitive to implicit dynamic information even when they are not able to observe real-time change
- perception requires information about *transitions*
- *dynamic* mental representations: temporal dimension is necessary part of it
- representational momentum
- dynamic information about possible past and future $\rightarrow$ no explicit mention of counterfactuals