Faulty Towers: A hypothetical 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

In this paper we introduce the hypothetical simulation model (HSM) of physical support. The HSM predicts that people judge physical support by mentally simulating what would happen if the object of interest were removed. Two experiments test the model by asking participants to evaluate the extent to which one brick in a tower is responsible for the rest of the bricks staying on a table. The results of both experiments show a very close correspondence between hypothetical simulations and responsibility judgments. We compare three versions of the HSM which differ in how they model people's uncertainty about what would happen. Participants' selections of which bricks would fall are best explained by assuming that hypothetical interventions only lead to local changes while leaving the rest of the scene unchanged.

Keywords: causality; counterfactual; hypothetical; mental simulation; intuitive physics; physical support.

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 different towers are and what is causing that stability (Battaglia, Hamrick, & Tenenbaum, 2013; Hamrick, Battaglia, Griffiths, & Tenenbaum, 2016). In this paper, we look at how people judge the extent to which different bricks carry the responsibility for a tower's stability. We argue that people judge responsibility by imagining what would happen to the tower if the brick were removed, and develop a *hypothetical simulation model* (HSM) of physical support which captures this process.

We build on previous work in which we have shown how a *counterfactual simulation model* (CSM) 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 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. 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 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, in press).

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 a model of hypothetical simulation naturally handles judgments about physical support.

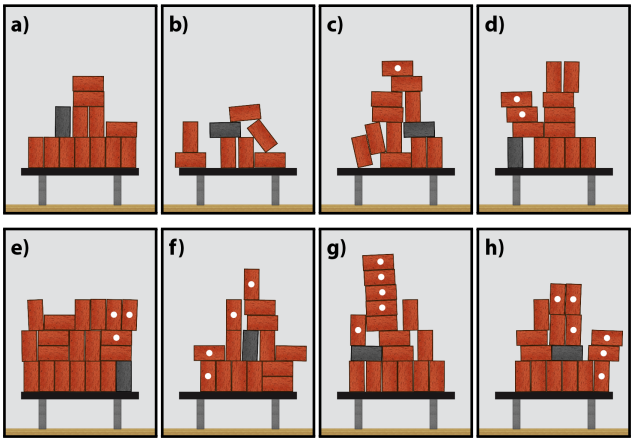
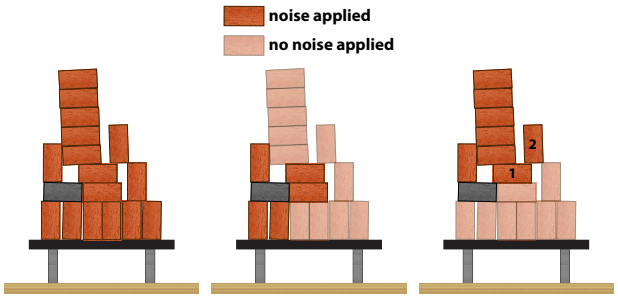


Figure 1: **Experiment 1**. Example stimuli. *Note*: Red bricks that would fall off the table if the black brick were removed (according to ground truth) are marked with a white dot at their center. The dots were not displayed in the actual experiment.

Judging physical support is different from judging causation in several ways. First, hypotheticals are different from counterfactuals in that they are future-oriented and don't require going back in time (Beck, 2015). When making causal judgments about dynamic collisions, the observer needs to remember what actually happened, and contrast this with what would have happened in the relevant counterfactual situation. However, when making judgments of physical support in static scenes, like the tower configurations in Figure 1, there is no need to go back in time. We merely need to simulate what a possible future would look like in which certain aspects of the scene were changed.

Second, the mental simulations that are required to imagine the relevant counterfactual or hypothetical are different (cf. Freyd, Pantzer, & Cheng, 1988; Holmes & Wolff, 2010). When simulating counterfactuals, we want to stay as close as possible to what actually happened, and only modify the world as little as possible to make the counterfactual true (Gerstenberg, Bechlivanidis, & Lagnado, 2013; Lewis, 1973; Pearl, 2000). But what do we keep constant in the causal model of the situation, and what do we change? When judging whether a ball would have gone into the goal, we need to simulate what the trajectory of the ball would have been if the collision hadn't taken place. To model people's uncertainty, we can add noise to the simulation of the ball's trajectory (cf. Smith & Vul, 2013) and keep everything else that we know about the scene as it was (e.g. we wouldn't change the size



(a) global noise (b) local noise (c) above noise

Figure 2: Schematic illustration of how different versions of the *hypothetical simulation model* apply noise when considering what would happen if the black brick were removed.

of the goal in the counterfactual simulation). However, when judging responsibility for a tower's stability, it is less clear what aspects of the scene we should hold constant. We will compare several implementations of the HSM that differ in how they capture people's uncertainty about what would happen.

The road map for the rest of the paper is as follows: We first present in detail how the HSM predicts judgments of physical support. We will test the model in two experiments in which we ask one group of participants to make hypothetical judgments, and another to evaluate causal responsibility. As predicted by the HSM, there is a very close correspondence between hypothetical and responsibility judgments. Heuristic strategies that focus on features of the scene (such as a tower's height, or the number of bricks on top of the brick of interest) cannot explain people's judgments as well. We end by discussing limitations of the current approach and by offering directions for future research.

Hypothetical simulation 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 HSM we need to determine (1) what hypothetical situation to consider, and (2) how to simulate what would happen in that situation. We assume that when judging responsibility, participants consider a hypothetical situation in which the black brick is removed. Participants then use their intuitive understanding of physics to mentally simulate what would happen in that situation.

Recent work has argued that some aspects of people's intuitive understanding of physics are well-described by assuming we have an approximate simulation engine in our mind that is akin to a physics engine (Battaglia et al., 2013; Lake, Ullman, Tenenbaum, & Gershman, 2016). Part of what makes these simulation engines "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).

To investigate whether people's mental simulations incorporate the assumption that only some aspects of the physical scene would directly be affected by the hypothetical intervention, we contrast three implementations of the HSM. These implementations differ in how they capture people's uncertainty about what would happen if the black brick were removed. All models apply noise in the same way: as a small impulse to some of the red bricks immediately after the removal of the black brick. The models differ, however, in which bricks they apply noise to. Figure 2 illustrates how the three different models work. The *global noise* model applies a small impulse to all the bricks and thus captures a general uncertainty about the scene (cf. Battaglia et al., 2013). The *local noise* model applies the impulse only to the red bricks that are directly in contact with the black brick. This model captures the assumption that participants will be most uncertain about what would happen in the area around the black brick. The *above noise* model applies noise only to bricks that are above the black brick and "connected" with it. Any brick that directly contacts and has its center of mass above that of the black brick counts as connected. This definition is then applied recursively. For example, brick 2 in Figure 2c is connected since brick 1 is in contact with and above the black brick, and brick 2 is in contact and above brick 1. This model captures that removing the black brick will affect the other bricks in an asymmetric way. Similar to when we lift a wooden block playing Jenga, this version of the model assumes that we have uncertainty particularly about those parts of the scene that would be affected by this kind of manipulation.

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 weren't there, or evaluate the extent to which the black brick is responsible for the red bricks staying on the table. In line with the HSM, we predicted that there would be a close relationship between hypothetical and responsibility judgments.

Methods

Design & Procedure The experiment had three conditions that differed only in terms of the dependent measure.[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 *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 *responsibility condition*, participants were asked to answer the question: "How responsible

[1]Data, materials, figures, and code are available here: https://github.com/tobiasgerstenberg/tower_counterfactual
An interface to view the stimuli and play around with the different noise models may be accessed here: http://web.mit.edu/tger/www/demos/towers/physics_interface.html

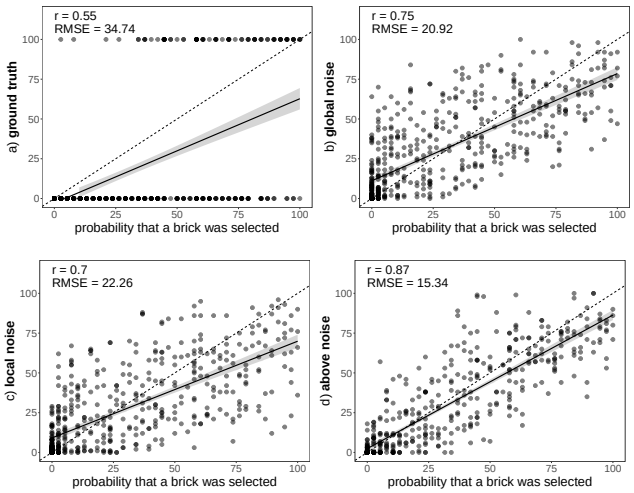


Figure 3: **Experiment 1**. Scatter plots showing the relationship between the empirical probability with which each brick was selected and (a) the ground truth as well as the predictions of the best-fitting (b) global noise model, (c) local noise model, and (d) above noise model.

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 warm-up animations that showed 20 bricks being dropped on the table. These animations were shown to familiarize participants with the relevant properties of the physical scene such as gravity, the friction between the bricks, as well as the table friction. Participants were only allowed to proceed to the next stage once they had watched at least five animations.

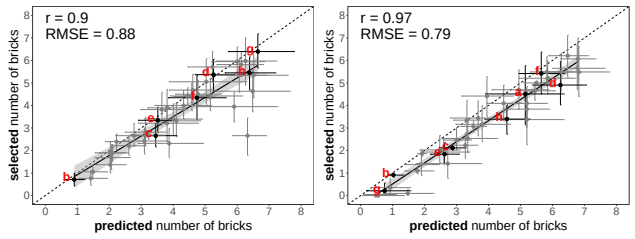
After the warm-up, participants saw 42 images of different towers of bricks in randomized order (see Figure 1 for examples). The stimuli varied the number of bricks on the table (range = 7 to 20, M = 13.7, SD = 3.3), as well as the number of red bricks that would fall off the table if the black brick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

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

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 that is applied to different bricks depending on the model.



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

Figure 4: Relationship between the predicted number of red bricks that would fall if the black brick weren'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 in all figures denote bootstrapped 95% confidence intervals.

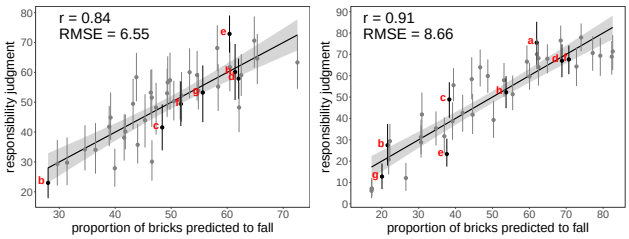
Participants 121 participants (M_{age} = 34, SD_{age} = 12, 47 female) were recruited via Amazon Mechanical Turk using psiTurk (Gureckis et al., 2016) with $N = 38$ in the selection condition, $N = 42$ in the prediction condition, and $N = 41$ in the responsibility condition. We excluded participants from further analysis based on their responses to the catch trial shown in Figure 1a. Eleven participants in the prediction condition were excluded because they predicted that at least one red brick would fall. Six participants in the responsibility condition were excluded because they gave a responsibility rating greater than 15. No participants were excluded from the selection condition because no participant selected any of the bricks on the catch trial.

Results

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

Selection condition We tested how well the three different noise models captured participants' selections of which bricks would fall off the table if the black brick weren't there (see Figure 2). For each model, we used maximum likelihood fitting to find 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. (Figure 8 gives an example for what these predictions look like for stimuli used in Experiment 2.) Overall, the *above noise* model accounted best for the data (cf. Table 1).

Prediction condition Figure 4a 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' hypothetical 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 $r = .88, \text{RMSE} = 0.84$.



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

Figure 5: Relationship between the predicted proportion of bricks that would fall if the black brick weren'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.

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 weren't present in the scene, and participants' responsibility judgments. As predicted by the HSM, there was a very close relationship between prediction and responsibility judgments $r = .84$, RMSE $= 6.55$. This suggests that participants evaluated a brick's responsibility by considering what proportion of bricks would fall off the table if the brick weren't 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 $= 7.65$. A noise-free model that uses the proportion of bricks that actually fall off the table does not account well for participants' responsibility judgments $r = .35$, RMSE $= 11.42$.

As an alternative to the HSM, we compared a heuristic model which predicts participants' responsibility judgments based on features of the physical scene. Table 2 shows how well the different features correlated with participants' judgments individually, as well as when combined via a linear regression model. We included features about the whole scene such as the number of bricks, the tower height, the average distance of each brick to the nearest edge of the table, as well as the average height and angle of each brick. We also included features specific to the black brick such as its distance to the nearest edge, its height and angle, as well as the number of bricks above it. To define the number of bricks above, we used the same criterion as the *above noise model* (cf. Figure 2c). As Table 2 shows, the best individual predictor for participants' responsibility judgments is the average height of each brick in the scene, followed by the number of bricks above the black one. Neither individual feature describes participants' responsibility judgments as well as the predictions

Table 2: Correlation coefficients between features and participants' responsibility judgments in Experiments 1 and 2. *Note*: The *scene features*, *brick features*, and *all features* columns show how well regressions that combine these features correlate with participants' judgments.

	scene features						**black brick features**					
	n bricks	tower height	avg edge distance	avg height	avg angle	scene features	edge distance	height	angle	n bricks above	brick features	all features
Exp 1	.16	.55	.39	.73	.21	.81	.02	-.19	-.05	.61	.62	.88
Exp 2	-.05	.21	-.10	.07	.01	.26	.12	-.74	-.04	.69	.79	.85

Note: n = number, avg = average.

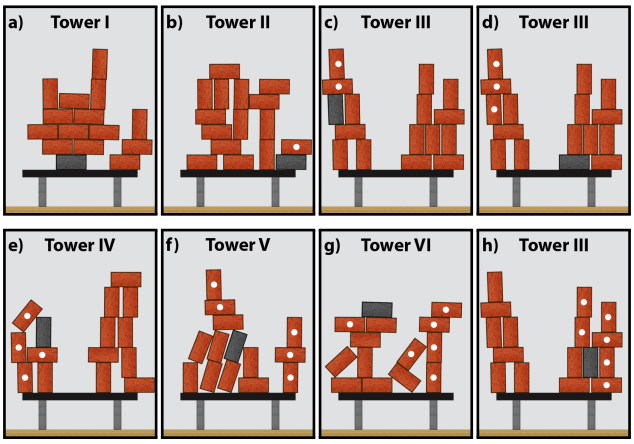


Figure 6: **Experiment 2**. Example stimuli. *Note*: White dots indicate which bricks would fall if the black brick weren't there. There were 6 different configurations of towers (I through VI), and 7 different positions for the black brick in each tower, see c), d), and h).

(and selections) that participants made in the other two conditions of the experiment. A regression model that combines all features correlates well with participants' responsibility judgments ($r = .88, \text{RMSE} = 5.89$), as does a model that only considers the scene features ($r = .81, \text{RMSE} = 7.14$). A model which only includes features about the black brick doesn't fare as well ($r = .62, \text{RMSE} = 9.6$). Even though a model that includes all features explains slighlty more of the variance that the HSM, this is likely due to overfitting; using model selection criteria, we find that the HSM performs better (AIC $= 276.52$, BIC $= 281.66$) than the heuristic model (AIC $= 283.72$, BIC $= 302.57$).

Discussion

The results of Experiment 1 support the predictions of the HSM. 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 weren't there. A heuristic model that does not rely on physical simulations but uses features that can be directly extracted from the scene fared worse when taking into account both variance explained and model complexity. We contrasted three implementations of the HSM which differ in the way in which they capture people's uncertainty about what would happen if the brick were removed. The results show that the *above noise* model correlates best with participants' selections. This model assumes that participants are particularly uncertain about what would happen to the bricks that are located above the black one.

Experiment 2

Experiment 1 elicited participants' judgments for a wide array of different situations. In Experiment 2, we chose a more tightly controlled stimuli set, a selection of which is shown in Figure 6. We generated six different tower configurations. For each configuration, we chose seven positions for the black brick such that removing it would result in 0 to 6 red bricks falling off the table. While a heuristic model that used

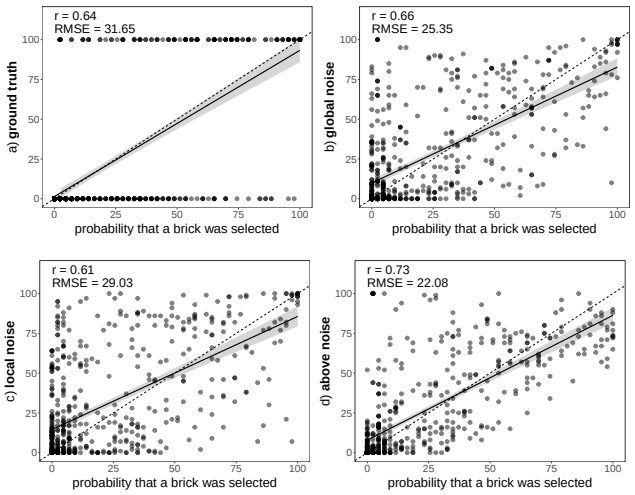


Figure 7: **Experiment 2**: Scatter plots showing the relationship between the empirical probability with which each brick was selected and (a) the ground truth as well as the predictions of the best-fitting (b) global noise model, (c) local noise model, and (d) above noise model.

global scene features explained responsibility judgments well in Experiment 1, we expected this model to perform poorly here since it doesn't take into account where the black brick is positioned.

In order to better tell apart the different implementations of the HSM, we included tower configurations with disjointed sets of bricks (Tower III and Tower IV). For example, consider the configuration of bricks shown in Figure 6c. While a global noise model predicts that some of the red bricks on the right would fall off the table, the local versions of the model predict that only the bricks on the left side will fall.

Methods

Design & Procedure The design, procedure, and questions were identical to those of Experiment 1. Participants saw 43 trials in randomized order whereby one trial served as a catch trial. On average, the experiment took 13.04 (SD = 6.87), 11.57 (SD = 5.24) and 7.86 minutes (SD = 3.48) in the selection, prediction, and responsibility condition, respectively.

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. We used the same exclusion criteria as in Experiment 1 based on the same tower shown in Figure 1a. 1 participant was removed in the selection condition, 3 participants in the prediction condition, and 3 in the responsibility condition.

Results & Discussion

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. Overall, the *above noise* model accounted best for participants' selections, as in Experiment 1 (cf. Table 1).

Let us look at the two situations shown in Figure 8 in some

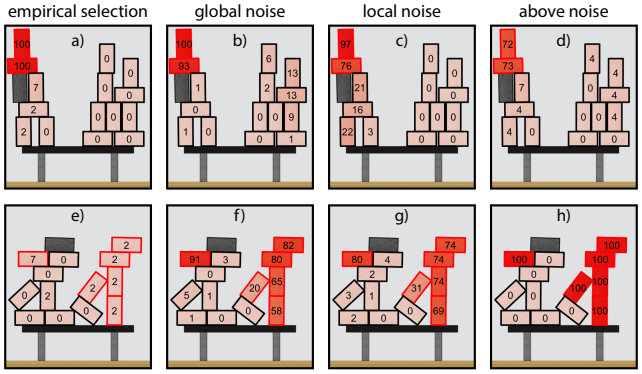


Figure 8: Empirical selection percentages for two different stimuli together with the predicted selection probabilities according to the different noise models. The numbers (and color fill) indicate what percentage predicted that a particular brick would fall off the table if the black brick were removed. Red and black frames around a brick indicate that the brick would fall or stay on the table, respectively.

more detail. For the example shown in the top row, participants' selections corresponded closely to the ground truth. Since the *global noise* model applies an impulse to all the bricks, it incorrectly predicted that participants would select bricks on the right. The *local noise* model incorrectly predicted selections of bricks underneath the black one. The *above noise* model best predicted participants' selection in this case. It only assigned a small probability that any of the bricks on the right would be selected (because sometimes the bricks on top of the black brick will fall towards the right), or bricks that are underneath the black one.

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 weren't there. 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 was able to capture participants' selections in this case. The *above noise* model did 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' hypothetical simulations. We will discuss some ideas about how the improve the models in the General Discussion below.

Prediction condition Figure 4b shows the relationship between the number of bricks predicted to fall and the average number of bricks that participants selected in the selection condition. As in Experiment 1, there was a very close relationship between predictions and selections, and, again, participants predicted that more bricks would fall on average than they selected. The *above noise* model again best explained participants' predictions with $r = .76, \text{RMSE} = 1.41$.

Responsibility condition Figure 5b shows the relationship between participants' predictions and responsibility judgments. Like in Experiment 1, participants' responsibility judgments were well-accounted for by the proportion of

bricks that would fall off the table if the black brick were removed $r = .91, \text{RMSE} = 8.66$. Again, we can also account for participants' responsibility judgments based on the proportion of bricks that were selected in the selection condition $r = .91, \text{RMSE} = 8.67$. A noise-free model again fails to account well for participants' responsibility judgments with $r = .36, \text{RMSE} = 19.99$.

Table 2 shows how well different features of the physical scene correlate with participants' responsibility judgments in Experiment 2. Expectedly, global scene features did not correlate well with participants' responsibility judgments this time because these features do not capture the actual position of the black brick. For example, they don't distinguish the configuration shown in Figure 6c from the one shown in Figure 6h. However, a good predictor of participants' responsibility judgments was the height of the black brick. The lower the black brick was located, the more responsible it was. Unlike in Experiment 1, the average height of the bricks in the tower did not correlate with responsibility judgments. Unsurprisingly, the number of bricks above the black brick was again a good predictor. However, there was no single predictor that accounted as well for participants' responsibility judgments as participants' predictions or selections in the other two conditions did. Even a regression that combines both scene and black brick features ($r = .85, \text{RMSE} = 11.17$) does not explain participants' responsibility judgments as well as the HSM does.

General Discussion

How do people judge physical support? In this paper, we develop and test a hypothetical simulation model (HSM) of physical support. Based on a model of counterfactual simulation which was originally developed to explain causal judgments about collision events (Gerstenberg et al., 2012, 2014, 2015; Gerstenberg & Tenenbaum, 2016), the HSM predicts that we judge physical support by imagining what would happen if the object were removed. An individual brick is responsible for other bricks staying on a table to the extent that these bricks would fall off the table if that brick were removed. The results of two experiments show that the greater the proportion of bricks that participants predict would fall of the table, the more responsible that brick is seen for the other bricks staying on the table. Simple features of the physical scene such as the height of the tower, or the position of the brick of interest, as well as combinations of these features cannot explain participants' judgments as well.

The central claim of the HSM is that people judge physical support by simulating what would happen to the scene if the object of interest were removed. We contrasted three different implementations of the HSM which differ in how they model participants' uncertainty about what would happen in the relevant hypothetical situation. Similar to how people spontaneously consider counterfactuals when judging causation (Gerstenberg et al., in press), people naturally play "mental Jenga" when judging responsibility for physical support. Participants' selections of which bricks would fall were best explained by a model that adds noise to the bricks located above the removed brick. While this model does a good job overall, there remain situations that it cannot capture adequately (cf. Figure 8).

We believe that there are at least three sources of uncertainty that affect participants' judgments: first, there is perceptual uncertainty about the exact spatial location of the different bricks (cf. Battaglia et al., 2013). Second, there is uncertainty about the hypothetical intervention itself: would the black brick simply disappear, or would it be removed, thereby affecting the bricks above it. Third, there is dynamic uncertainty about what would happen once the brick is removed (cf. Smith & Vul, 2013). While the current implementation of the HSM uses a physics engine as a proxy for participants' mental model, we are eager to explore how an approximate simulation model (which doesn't represent each brick individually) might be able to capture participants' judgments (cf. Davis & Marcus, 2016). Ideally, such a model would help explain when it is that people's physical intuitions are faulty and deviate from the ground truth.

Acknowledgments We thank Sunny Khemlani and two anonymous reviewers for their comments on the paper. 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.

Beck, S. R. (2015). Why what is counterfactual really matters: A response to Weisberg and Gopnik (2013). *Cognitive Science*, *40*(1), 253–256.

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

Freyd, J. J., Pantzer, T. M., & Cheng, J. L. (1988). Representing statics as forces in equilibrium. *Journal of Experimental Psychology: General*, *117*(4), 395–407.

Gerstenberg, T., Bechlivanidis, C., & Lagnado, D. A. (2013). Back on track: Backtracking in counterfactual reasoning. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Conference of the Cognitive Science Society* (pp. 2386–2391). Austin, TX: Cognitive Science Society.

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. (in press). Eye-tracking causality. *Psychological Science*.

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.

Gureckis, T. M., Martin, J., McDonnell, J., Rich, A. S., Markant, D., Coenen, A., … Chan, P. (2016). psiturk: An open-source framework for conducting replicable behavioral experiments online. *Behavior Research Methods*, *48*(3), 829–842.

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.

Holmes, K. J., & Wolff, P. (2010). Simulation from schematics: dorsal stream processing and the perception of implied motion. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 2704–2709). Austin, TX: Cognitive Science Society.

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*.

Lewis, D. (1973). Causation. *The Journal of Philosophy*, *70*(17), 556–567.

Pearl, J. (2000). *Causality: Models, reasoning and inference*. Cambridge, England: Cambridge 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.