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

In this paper we extend the counterfactual simulation model (CSM) – originally developed to capture causal judgments about dynamic events (Gerstenberg, Goodman, Lagnado, & Tenenbaum, 2014) – to explain judgments of physical support. The CSM 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 the table. The results of both experiments show a very close correspondence between counterfactual simulations and responsibility judgments. We compare three versions of the CSM which differ in how they model people's uncertainty about what would have happened. Participants' selections of which bricks would fall are best explained by assuming that counterfactual interventions only affect some aspects while leaving the rest of the scene unchanged.

Keywords: causality; counterfactual; mental simulation; intuitive physics; 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 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). Moreover, we can also point to *why* the tower is stable. We can judge the extent to which different bricks carry the responsibility for the tower's stability. In this paper, we propose how people do so: when judging responsibility, people imagine what would happen to the tower if the brick were removed. The greater the proportion of bricks that would fall off the table in their mental simulation, the more responsible is the brick of interest. We develop a *counterfactual simulation model* (CSM) of physical support which determines a brick's causal responsibility for the tower's stability by simulating scenarios that determine what would happen if the brick were removed. Supporting doesn't simply mean "being underneath", it means "preventing from falling".

In previous work, we showed how the CSM explains people's causal judgments about dynamic collision events (Gerstenberg, Goodman, Lagnado, & Tenenbaum, 2012; Gerstenberg et al., 2014; Gerstenberg, Goodman, Lagnado, & Tenenbaum, 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, submitted). 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 as well.

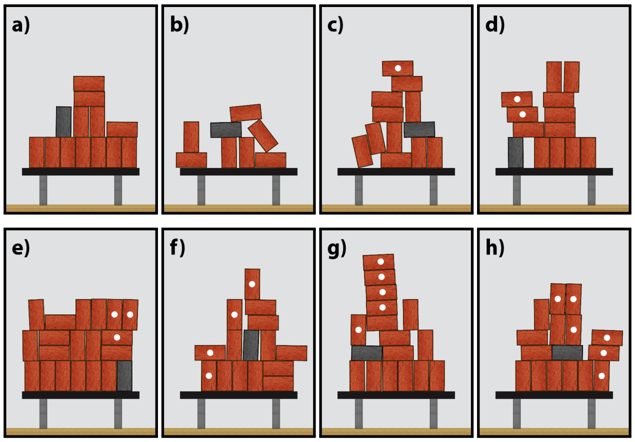


Figure 1: **Experiment 1**. Example stimuli. *Note*: Red bricks that would fall off the table if the black brick were removed (according to ground truth) are marked with a white dot at their center. The dots were not displayed in the actual experiment.

Judging 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). For instance, it is a player's kick that causes a ball to go into a goal, or a collision event between two balls that causes one of them to go through a gate. 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 (cf. Freyd, Pantzer, &

Cheng, 1988; Holmes & Wolff, 2010). So, rather than defining a counterfactual operation on events, the CSM considers counterfactual operations on objects in the scene. The more certain we are that the tower would collapse if the brick were removed, the more responsible it is for the tower's stability.

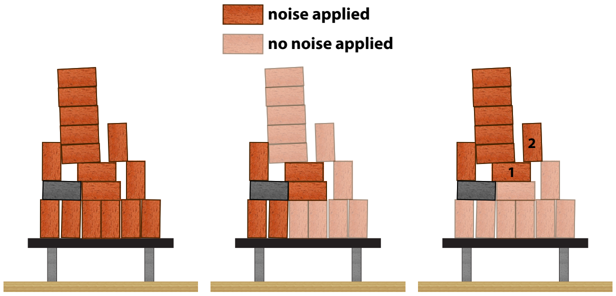
Another important point is the role that uncertainty plays in people's counterfactual simulations. When simulating counterfactuals, we want to stay as close as possible to what actually happened, and only modify the world as little as possible to make the counterfactual true (Gerstenberg, Bechlivanidis, & 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 simulating counterfactuals? 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 would add noise to the ball's trajectory (cf. Smith & Vul, 2013) but keep everything else that we know about the scene as it was (e.g. we wouldn't change the size 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 CSM that differ in how they capture people's uncertainty about what would have happened.

The road map for the rest of the paper is as follows: We first present in detail how the CSM predicts judgments of physical support. We will test the model in two experiments in which we ask one group of participants to make counterfactual judgments, and another to evaluate causal responsibility. As predicted by the CSM, there is a very close correspondence between counterfactual 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.

Counterfactual 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 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 were removed. Participants then use their intuitive understanding of physics to mentally simulate what would happen in that situation.[1]

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).



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

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.

To investigate whether people's mental simulations incorporate the assumption that only some aspects of the physical scene would directly be affected by the counterfactual intervention, we contrast three implementations of the CSM. These implementations differ in 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 some of the red bricks immediately after the removal of the black brick. The impulse is applied in a random direction, with the amplitude determined by sampling from a Gaussian parametrized by a "noise level". 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. 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. Any brick that directly contacts and is has 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.

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.

[1] We use the term *counterfactual* broadly here to refer to simulations of possible worlds. Judging causation for dynamic collision events requires the observer to remember what happened, and go back in time to construe the counterfactual. When judging physical support, there is no need to go back in time. A *hypothetical* simulation of what would happen in the future suffices.

Methods

Design & Procedure The experiment had three conditions that differed only in terms of the dependent measure.[2] 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 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 the gravity, 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 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 give 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.

Participants 121 participants (M_{age} = 34, SD_{age} = 12, 47 female) were recruited via Amazon Mechanical Turk 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. 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. 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.

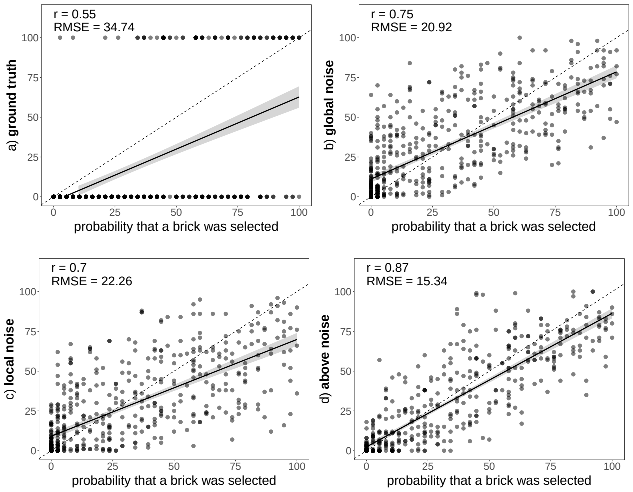


Figure 3: **Experiment 1**. Scatter plots showing the relationship between the empirical probability with which each brick was selected and the (a) ground truth as well as the predictions of the best-fitting (b) global noise model, (c) local noise model, and (d) above noise model.

Selection condition 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 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 accounts best for the data (cf. Table 1).

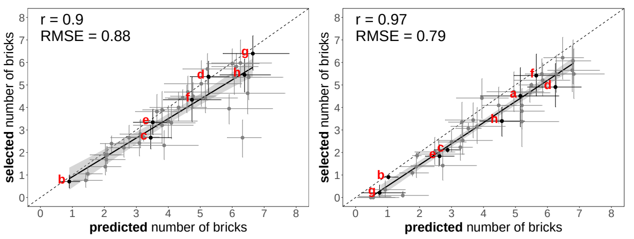
Prediction condition Figure 4 shows the relationship between the number of bricks predicted to fall and the average number of bricks that participants selected in the selection condition. Overall, the two ways of probing participants' counterfactual simulations lead to very similar results. However, 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 $r = .88, \text{RMSE} = 0.84$.

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] 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`



(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 in all figures denote bootstrapped 95% confidence intervals.

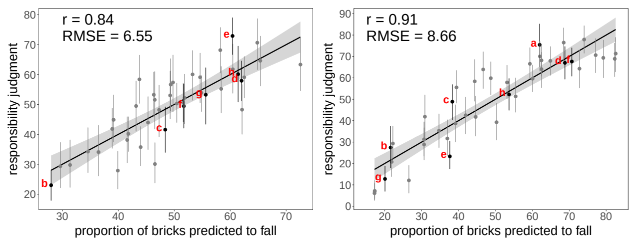
Responsibility condition Figure 5a shows the relationship between the proportion of bricks that participants in the *prediction condition* believed would fall off the table if the black brick wasn't present in the scene, and participants' responsibility judgments. As predicted by the CSM, there was a very close relationship between prediction and responsibility judgments $r = .84$, RMSE $= 6.55$. 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 $= 7.65$.

As an alternative to the CSM, 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 individually correlated with participants' judgments. 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 feature describes participants' responsibility judgments as well as the predictions (and selections) participants made in the other two conditions of the experiment.

Table 2: Correlation coefficients between different features and participants' responsibility judgments in Experiments 1 and 2.

	scene features					**black brick features**			
	n bricks	tower height	avg edge distance	avg height	avg angle	edge distance	height	angle	n bricks above
Experiment 1	.16	.55	.39	.73	.21	.02	-.19	-.05	.61
Experiment 2	-.05	.21	-.10	.07	.01	.12	-.74	-.04	.69

Note: n = number, avg = average.



(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.

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. A heuristic model that does not rely on physical simulations but uses features that can be directly extracted from the scene did not find a single predictor that accounted as well for participants' responsibility judgments as participants' predictions did. We contrasted three implementations of the CSM 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 model* correlates best with participants' selections. It is thus not surprising that the the number of bricks above the black brick (used as a feature) correlates well with participants' responsibility judgments.

Overall, participants' responsibility judgments were slightly better explained by participants' judgments in the prediction condition than the selections in the selection condition. While predictions and selections were highly correlated, participants tended to predict that more bricks would fall on average than they selected. The time it took participants to complete the experiment in the different conditions suggests that participants in the selection condition may have engaged more deeply with the stimuli than participants in the other two conditions did.

Experiment 2

Experiment 1 elicited participants' judgments for a wide array of different situations. In Experiment 2, we wanted to test the different implementations of the CSM in a more controlled setup. Figure 6 shows a selection of the stimuli. Some of the configurations featured disjointed sets of bricks, such as Tower III and Tower IV, for which the predictions of the global and local noise models differ more strongly. 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. We generated six different tower configurations. For each configuration, we chose seven positions for the black brick such that removing the brick would result in 0 to 6 red bricks falling off the table. Figure 6c, d, and g show three

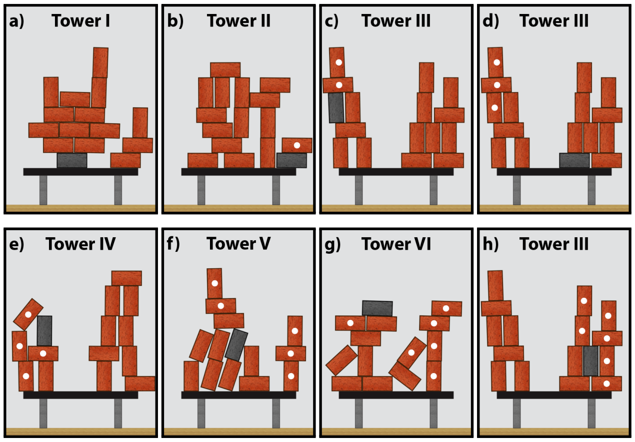


Figure 6: **Experiment 2**. Example stimuli. *Note*: The white dots indicate which bricks would fall if the black brick wasn't there. There were 6 different configurations of towers (I through VI), and 7 different positions for the black brick in each tower, see c), d), and h).

examples for the position of the black brick for this configuration. This means that global features of the scene will not be diagnostic for how many bricks would fall off the table if the black brick were removed (cf. Table 2).

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 participants were 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. Across all the stimuli, there were 564 bricks in total. Overall, the *above noise* model accounts for participants' selections best, as in Experiment 1 (cf. Table 1).

Let us look at 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.

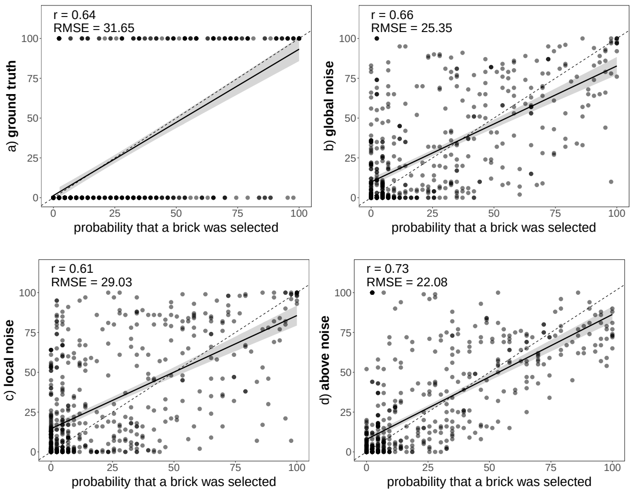
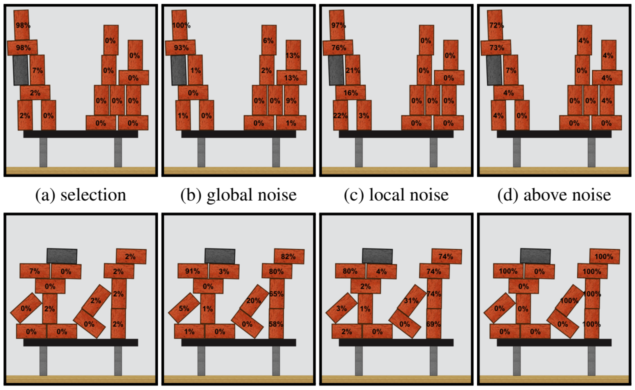


Figure 7: **Experiment 2**: Scatter plots showing the relationship between the empirical probability with which each brick was selected and the (a) ground truth as well as the predictions of the best-fitting (b) global noise model, (c) local noise model, and (d) above noise model.

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 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 $r = .76, \mathrm{RMSE} = 1.41$.

Responsibility condition Figure 5b shows the relationship between participants' predictions and responsibility judg-



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

(e) selection (f) global noise (g) local noise (h) above noise

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

ments. 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 $r = .91, \mathrm{RMSE} = 8.66$. Again, we can also account for participants' responsibility judgments based on the proportion of bricks that were selected in the selection condition $r = .91, \mathrm{RMSE} = 8.67$.

Table 2 shows how well different features of the physical scene correlate with participants' responsibility judgments in Experiment 2. This time, 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, again, there was no single predictor that accounted as well for participants' responsibility judgments as the predictions or selections did.

General Discussion

How do people judge physical support? In this paper, we develop and test a counterfactual simulation model (CSM) of physical support. The CSM predicts that we judge physical support by imagining what would happen if the object were removed. An individual brick is responsible for other brick's 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, cannot explain participants' judgments as well.

While the CSM was originally developed to explain causal judgments about collision events (Gerstenberg et al., 2012, 2014, 2015; Gerstenberg & Tenenbaum, 2016), we show here that it naturally extends to judgments of physical support. Indeed, the results here highlight the importance of having a model that is flexible in how counterfactual interventions are defined, and their consequences simulated. For explaining causal judgments, we need to consider counterfactuals over events (such as collisions). For explaining support judgments, we need to consider counterfactuals over objects, and simulate what would happen if they were removed. We contrasted three different implementations of the CSM which differ in how they modeled participants' uncertainty about what would happen. Similar to how people spontaneously consider counterfactuals when judging causation (Gerstenberg et al., submitted), 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, and we will explore more fully in future work how people simulate the consequences of counterfactual interventions.

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., 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. (submitted). 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.

Holmes, K. J., & Wolff, P. (2010). Simulation from schematics: dorsal stream processing and the perception of implied motion. In *Proceedings of the 32nd annual conference of the cognitive science society* (pp. 2704–2709).

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.

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

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.