Mental Jenga
A counterfactual simulation model of physical support

Liang Zhou
University College London

Kevin A. Smith
Massachusetts Institute of Technology

Joshua B. Tenenbaum
Massachusetts Institute of Technology

Tobias Gerstenberg[*]
Stanford University

Abstract

From building towers to picking an orange from a stack of fruit, assessing support is critical for successfully interacting with the physical world. But how do people determine whether one object supports another? In this paper we develop the Counterfactual Simulation Model (CSM) of physical support. The CSM predicts that people judge physical support by mentally simulating what would happen to a scene if the object of interest were removed. Three experiments test the model by asking one group of participants to judge what would happen to a tower if one of the blocks were removed, and another group of participants how responsible that block was for the tower's stability. The CSM accurately captures participants' predictions about what would happen by running noisy simulations that incorporate different sources of uncertainty. Participants' responsibility judgments are closely related to counterfactual predictions: the more likely the tower would be predicted to fall if a block were removed, the more responsible this block was judged for the tower's stability. By construing physical support as preventing from falling, the CSM provides a unified account across dynamic and static physical scenes of how causal judgments arise from the process of counterfactual simulation.

Keywords: causality; counterfactual; responsibility; mental simulation; intuitive physics; physical support; sustaining causation.

[*] Corresponding author: Tobias Gerstenberg, Stanford University, Department of Psychology, 450 Jane Stanford Way, Bldg 420, Stanford, CA 94305, Email: gerstenberg@stanford.edu. All the data and study materials are available here: `https://github.com/cicl-stanford/mental_jenga`

Introduction

Take a look around yourself, and you'll notice something that's at the same time both perfectly ordinary and striking: most things don't move. The computer monitor doesn't move, the table on which it rests doesn't move, the floor on which the table stands doesn't move, and so on. Things don't move because they are supported by other things. The computer monitor is supported by the table, which is supported by the floor, which is supported by the structure of the house, which is supported by its foundation, and so on. But what does it mean for the monitor to be supported by the table? One intuitive answer is that the table supports the monitor because the monitor is *on* the table. But what if the monitor on the table was attached to a monitor arm that's drilled into the wall? Does the table still support the monitor in this case? Maybe support means something different.

In this paper, we explore the idea that people's understanding of physical support is intimately linked to their understanding of causation. One object A supports another object B if A prevents B from moving (or falling). What does it mean for A to prevent B from falling? The answer to this question involves a counterfactual: A prevents B from falling when it is true that B would fall if A were removed. We develop a computational model – the *counterfactual simulation model* (CSM) of physical support – that implements this idea and test the model in three sets of experiments asking participants to evaluate to what extent one object is responsible for the stability of other objects. We believe that people solve this task by constructing a mental model of the scene, and by simulating what would happen if the object of interest were removed.

Here is the road map for the paper: We first review prior work on people's intuitive understanding of the physical world, and on how people make causal judgments. We then describe the CSM in detail and contrast it with an alternative account that predicts people's judgments about physical stability based on surface-level features of the scene. We test the models in three experiments in which participants view towers of blocks that are stacked on a table. We ask one group of participants to judge what would happen if a block of interest were removed, and another group of participants how responsible that block is for all of the other blocks' staying on the table (Experiments 1 and 2), or for one specific other block (Experiment 3). Across these experiments, we find that the counterfactual predictions of one group of participants about what would happen if the block were removed are closely related to the responsibility judgments of another group of participants. We also find that the CSM accurately captures participants' judgments, and that a model that only uses surface-level features of the scene, such as the height of the tower and the location of the to-be-removed block, doesn't capture participants' judgments as well. We conclude by highlighting some limitations of the CSM, and some future challenges that lie ahead.

People's intuitive understanding of the physical world

People generally have a good sense for how the physical world works. We catch balls, stack stones, ride bikes, and build towers (Figure 1, top). While earlier work has documented how people's physical intuitions sometimes fail (McCloskey, 1983; McCloskey, Caramazza, & Green, 1980; McCloskey, Washburn, & Felch, 1983), more recent work has emphasized the successes (Kubricht, Holyoak, & Lu, 2017; Smith et al., under review). People can make predictions about the future such as whether a tower will topple over (Battaglia, Hamrick,

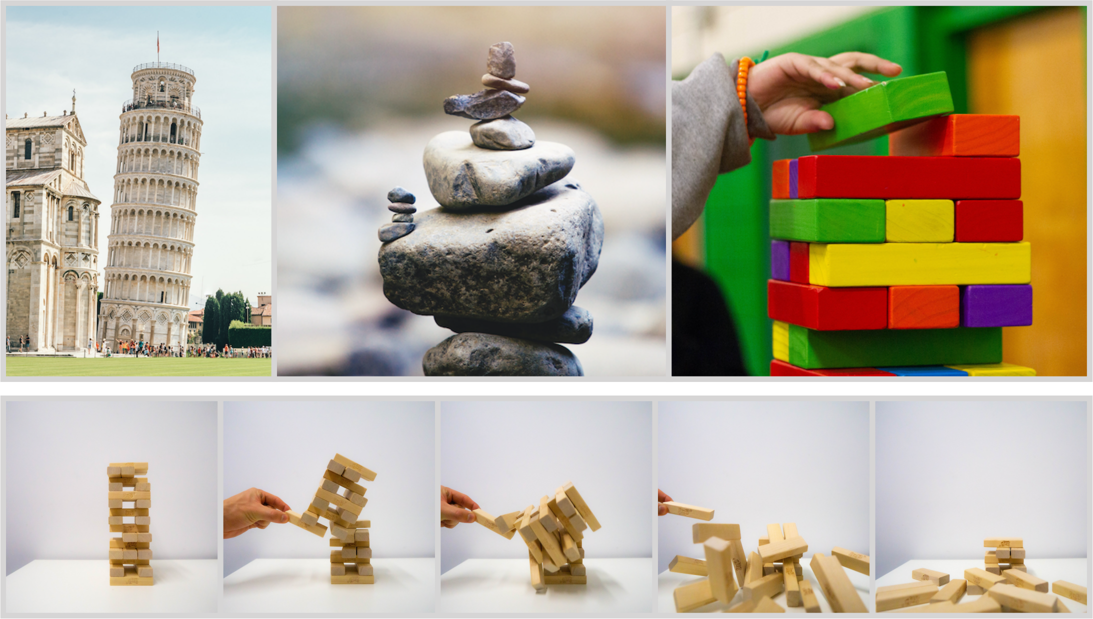


Figure 1. Stacked towers. **Top**: Real-life examples of towers, built from different materials. Left to right: Leaning tower of Pisa; a carefully balanced rock cairn; and a child stacking toy blocks. **Bottom**: We illustrate the process and consequences of removing a block from a tower in the *Jenga* game. Playing *Jenga* involves removing a block from a stable tower of blocks while attempting not to cause the tower to fall, like it does here.

& Tenenbaum, 2013), or where a moving object will go next (Smith, Dechter, Tenenbaum, & Vul, 2013; Smith & Vul, 2013). People make inferences about the past such as where a ball was behind an occluder before it appeared (Smith & Vul, 2014), in which hole of a box it was dropped based on where it landed and the sounds it made when it collided with the obstacles in the box (Gerstenberg, Siegel, & Tenenbaum, 2021), or whether a person used one or two hands to reconfigure a stack of blocks (Yildirim, Gerstenberg, Saeed, Toussant, & Tenenbaum, 2017; Yildirim et al., 2019). People can also infer unobservable physical properties such as the mass of different objects based on how they collided with one another (Sanborn, Mansinghka, & Griffiths, 2013), or based on the fact that a tower consisting of different objects is currently stable (Hamrick, Battaglia, Griffiths, & Tenenbaum, 2016). And people give causal explanations of what happened by comparing what actually happened with the counterfactual of what would have happened if the candidate cause hadn't been present (Gerstenberg, Goodman, Lagnado, & Tenenbaum, 2021; Gerstenberg & Icard, 2020; Gerstenberg, Peterson, Goodman, Lagnado, & Tenenbaum, 2017).

Underlying these various success stories is a common human feat: people reason about the physical world by building mental models (Craik, 1943; Gerstenberg & Tenenbaum, 2017; Sloman, 2005; Smith et al., under review; Ullman, Spelke, Battaglia, & Tenenbaum, 2017; Ullman & Tenenbaum, 2020). Prediction, inference, and explanation can be understood as different operations over these mental models. For example, Battaglia et al. (2013) model people's judgments about whether or not a tower of blocks is going to fall by assuming that people construct a mental model of the scene based on the perceived visual input, and then make predictions by mentally simulating how the physical scene will unfold (Schwartz & Black, 1999). While the physical world is deterministic – meaning there is a single true answer to the question of whether (and how) a tower will fall – people don't have access to this ground truth. Instead, they have to use their mental model to simulate what will happen. People's predictions about what will happen are graded: they don't know for sure whether or not a tower is going to fall.

Battaglia et al.'s model accurately captures the gradedness in people's responses by assuming that people are uncertain about different aspects of the scene, and that this uncertainty affects their mental simulations of what will happen. Specifically, the model assumes that people have some perceptual uncertainty about where exactly the different blocks are located, and some dynamic uncertainty about how exactly the scene is going to unfold. The model predicts people's judgments by starting with the actual configuration of the blocks, randomly perturbing the location of each one, and then simulating what will happen. To generate these simulations, the model uses the same physics engine that was used to make the stimuli. This process of random perturbation plus forward simulation is repeated multiple times to generate a distribution over possible future outcomes. This distribution is then used to capture people's graded judgments. For example, consider a stable tower A and another tower B that is on the brink of falling. Tower A is unlikely to fall even if each block's location was randomly perturbed. Tower B, however, is likely to fall when the block locations are perturbed. By taking the proportion of times in which a tower falls across the noisy simulations, the model yields a graded prediction about whether the tower will fall.

Block towers have emerged as somewhat of a drosophila for studying people's intuitive understanding of physics (Battaglia et al., 2013; Cortesa et al., 2018; Fischer, Mikhael, Tenenbaum, & Kanwisher, 2016; Gweon, Asaba, & Bennett-Pierre, 2017; Hamrick et al., 2016; Mitko & Fischer, 2020; Yildirim et al., 2017, 2019). Recent work has proposed different ways for how people might learn to make predictions about block towers and about other physical settings (Allen, Smith, & Tenenbaum, 2020; Baradel, Neverova, Mille, Mori, & Wolf, 2019; Battaglia, Pascanu, Lai, Rezende, et al., 2016; Bear et al., 2021; Bramley, Gerstenberg, Tenenbaum, & Gureckis, 2018; Chang, Ullman, Torralba, & Tenenbaum, 2017; Groth, Fuchs, Posner, & Vedaldi, 2018; Janner et al., 2019; Lerer, Gross, & Fergus, 2016; Ullman, Stuhlmüller, Goodman, & Tenenbaum, 2018; Wu, Yildirim, Lim, Freeman, & Tenenbaum, 2015). Our work builds on the idea that people's mental representation of the physical scene is in important ways similar to how the scene would be constructed in a physics engine of the kind that is used to make physically realistic animations in video games (Gerstenberg & Tenenbaum, 2017; Smith et al., under review; Ullman et al., 2017, but see Ludwin-Peery, Bramley, Davis, & Gureckis, 2021).

In the work presented here, we asked people to judge how responsible one block is for the stability of the tower. To answer this question, we need to turn to causality.

People's intuitive understanding of causality

People use their intuitive understanding of the physical world not only to make predictions about the future (e.g. where will this ball land?), but also to explain what happened (e.g. where did this ball come from, and who threw it?). While predictions and explanations may be operating on the same mental model, they require distinct computations (Gerstenberg & Tenenbaum, 2017). For prediction, one only needs to unroll a simulation of what will happen forward. Giving causal explanations, however, requires a comparison of what actually happened with what would have happened otherwise (Gerstenberg, 2022).

But are such counterfactual comparisons really necessary? Maybe it's sufficient to just focus on what can directly be perceived? In the philosophical literature on causation, there are two major families of theories for thinking about causation. According to process theories, causation is understood in terms of a transfer of a property via a spatiotemporally continuous process from cause to effect (Dowe, 2000; Salmon, 1994; Wolff, 2007). For example, A caused B to move when A transferred momentum to B. According to dependence theories, causation is understood as a form of dependence (Hume, 1748/1975; Lewis, 1973; Mackie, 1974; Suppes, 1970). For example, according to a counterfactual theory (Lewis, 1973; Pearl, 2000; Woodward, 2003), A caused B to move when B wouldn't have moved if A had been removed from the scene. Process theories ground causation merely in terms of what actually happened, whereas dependence theories involve a comparison between what actually happened and what could have happened otherwise.

Drawing on both theoretical frameworks, Gerstenberg, Goodman, et al. (2021) developed the *counterfactual simulation model* (CSM) of how people make causal judgments about dynamic physical events. In line with dependence theories, the CSM assumes that judging whether ball A caused ball B to go through a gate requires comparing what actually happened with what would have happened if ball A hadn't been present in the scene. In line with process theories, the CSM assumes that people's understanding of the underlying physical processes guides their mental simulations. Gerstenberg, Goodman, et al.'s (2021) experiments show that people's causal judgments are closely related to their beliefs about what would have happened in the relevant counterfactual situation. The more certain people are that the outcome would not have happened without ball A, the more they judge that ball A caused the outcome. Gerstenberg et al. (2017) showed that people spontaneously engage in counterfactual simulations when making causal judgments as evidenced by their eye-movements. People don't just look at what actually happened, they look toward where ball B would have gone if ball A hadn't been present in the scene (see also Gerstenberg, 2022).

Here, we build on this model and apply it to understanding people's judgments of physical support. Judging physical support is intimately related to judging causation. A supporting block is a sustaining cause of the tower's stability (Ross & Woodward, 2021). In other words, what it means for one object to support another is to prevent it from falling (or moving). If that's correct, then judging physical support requires going beyond what can be directly perceived, just like it does for judging causation. Judging physical support requires mentally simulating what would happen if the object of interest were removed. Judging causation and judging physical support thus rely on similar cognitive mechanisms. However, there are also a number of important differences.

First, the inputs differ. In Gerstenberg, Goodman, et al. (2021) the inputs to people's judgments of causation were short video clips depicting dynamic interactions. The inputs to people's judgments of support in the experiments reported here are images of static scenes (see Figure 4 for examples): they show block towers that are currently stable, but that might collapse if one (or more) of the blocks were removed. Because the scenes are static and stable without outside intervention, process theories of causation cannot capture the notion of physical support. Process theories rely on a transfer of a property, such as physical force, from one object to another (Wolff, 2007). While physical forces are clearly at play in keeping a block tower stable, they do not transfer between the objects as characterized by process theories. Moreover, theories of causation generally take the causal relata to be events: one event causes another event to happen (Schaffer, 2016). However, in the case of physical support, there aren't any events. The tower just stands still and nothing is moving. The notion of physical support thus broadens the concept of causation: causation doesn't only happen between events, it also "happens" when nothing is happening. The CSM provides a unifying framework for causal judgments across a variety of situations that span different types of causal relationships. It applies to "event causation" where a candidate cause event brings about some outcome event (Gerstenberg, Goodman, et al., 2021), to "omissive causation" where an effect comes about because a potential event didn't happen (Gerstenberg & Stephan, 2021), as well as to "sustaining causation" where the presence of the cause sustains the effect and where no events are happening at all.

Second, the different inputs require different kinds of computations (Beck, 2015). When judging whether ball A caused ball B to go through the gate, people have to go back in time to evaluate how things would have played out if ball A had been removed from the scene. When judging physical support, there is no need to go back in time because nothing is (or was) happening – the tower is just sitting still. Instead, one needs to imagine how a possible future might play out in which certain aspects of the scene were changed.[1]

Third, different mental simulations are required when making causal judgments about billiard balls versus block towers (cf. Freyd, Pantzer, & Cheng, 1988; Holmes & Wolff, 2010). When judging whether ball A caused ball B to go into the goal, people need to mentally simulate where ball B would have ended up if ball A hadn't been present in the scene. People may be uncertain about exactly what trajectory ball B would have taken in that counterfactual situation. When judging how responsible one block is for a tower's stability, people need to simulate what would happen to the tower if that block were removed. Again, there are multiple sources of uncertainty that may affect people's simulations including perceptual uncertainty about the position of each block, as well as dynamic uncertainty about how the collisions would play out. We will compare several implementations of the CSM that differ in how they capture people's uncertainty about what would happen.

In our experiments, we don't ask participants directly about physical support. Instead, we ask participants to judge the extent to which one block was responsible for other blocks (or one specific block) to stay on a table on which the blocks are stacked. Ques-

[1] When judging causation in dynamic scenes, people need to consider what *would have happened* if something in the past had been different, whereas when judging support in static scenes, people need to consider what *would happen* if something in the present were different. The question we ask participants in our experiments ("How many of the red bricks would fall off the table, if the black brick wasn't there?") sits right in the middle between a future-directed hypothetical question ("How many of the red bricks will fall off the table, if the black brick *isn't* there?") and a clearly counterfactual question ("How many of the red bricks would have fallen off the table, if the black brick *hadn't been* there?"). We believe that in our setting, participants would give the same response to any of these three versions of the question. Because of the static nature of the scene, counterfactual and hypothetical questions don't come apart. To highlight the continuity with our prior work (Gerstenberg, Goodman, et al., 2021), we chose to call our model the counterfactual simulation model of physical support, rather than the hypothetical simulation model.

tions of physical support may more naturally elicit binary responses (Does A support B?) whereas asking for responsibility elicits a graded response. That said, we believe that in our experiments, we would have observed very similar results if we had asked participants to judge physical support instead of (causal) responsibility. We have shown in previous work that in physical settings, responsibility judgments and causal judgments are closely related (see Gerstenberg, Goodman, et al., 2021). However, there are situations in which judgments about physical support and responsibility judgments could come apart. For example, if one block A lies on top of another block B that's positioned at the edge of a table, we might say that A prevents B from falling off the table (and is thus responsible for its staying on the table), but we would wouldn't say that A supports B. We return to the question of how exactly physical support and preventing from falling are related in the General Discussion.

To sum up, the main idea is that people judge physical support by considering whether the candidate object prevents the others from falling. Doing so requires mentally simulating what would happen if the object were removed. Judging physical support is like playing Jenga in your mind (Figure 1, bottom).

The Counterfactual Simulation Model (CSM)

The CSM predicts people's judgments about how responsible one object is for the stability of another object, or several other objects. At the core of the CSM is a noisy physics engine that supports interventions on the scene, such as removing an object, and running stochastic simulations of what would happen (Sanborn & Chater, 2016). If all the relevant physical parameters of the scene are fully specified (including the objects' mass and friction, their exact position and size, the elasticity that determines the "bounciness" of collisions, etc.), and if there is no ambiguity about what it means to remove an object from the scene (e.g. just making it disappear), then there is a deterministic ground truth answer as to what would happen. However, people don't have access to this ground truth. There are various sources of uncertainty that affect people's judgments (Battaglia et al., 2013; Smith & Vul, 2013). What these sources of uncertainty are will depend on the characteristics of the scene. In this paper, we consider block towers like the one shown in Figure 2. The question is to what extent the black block is responsible for the stability of the tower. Notice that our setup is different from Battaglia et al. (2013). In their experiments, participants viewed a static scene but "physics was turned off". The task was to predict what will happen once physics is turned back on again. In our setting, participants know that the tower currently is stable because physics is "on" and nothing is moving.

Sources of uncertainty in counterfactual simulation

We distinguish three sources of uncertainty: *perceptual uncertainty* about the position of each object, *intervention uncertainty* about how the black block is removed, and *dynamic uncertainty* about how the physical scene will unfold after the black block is removed.

Perceptual uncertainty. Participants know that the initial scene is stable. However, they may still be uncertain about the exact position of each block (Battaglia et al., 2013; Smith & Vul, 2013). Figure 2a illustrates how this is implemented in our model: it takes the ground truth configuration of blocks (shown faintly in the background) and then applies a small horizontal perturbation to each block's position, randomly moving some of

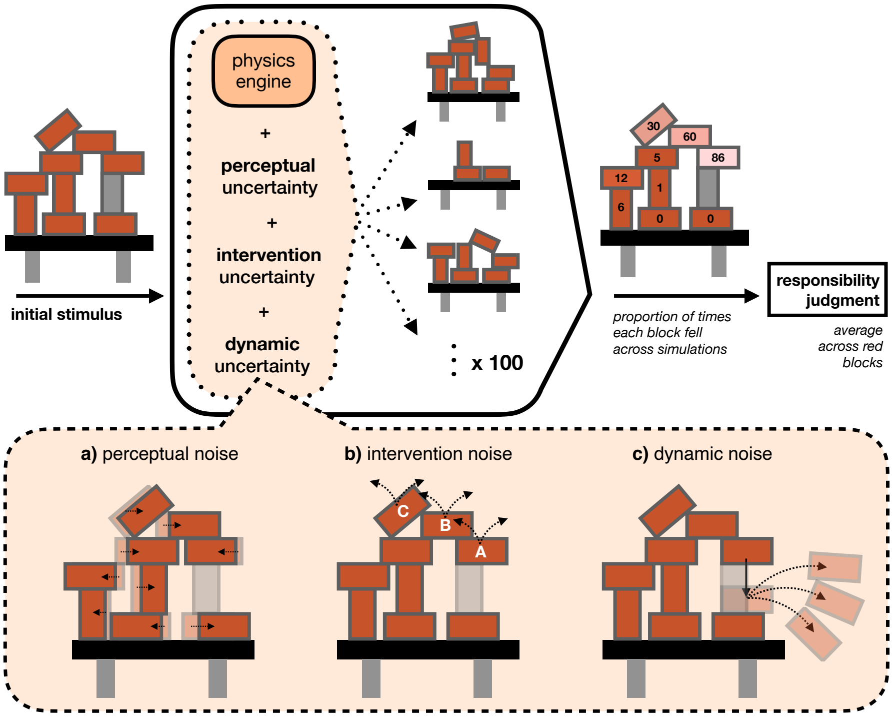


Figure 2. Schematic illustration of the *counterfactual simulation model* (CSM) applied to a block tower. Given a scene as input (top left), the CSM answers the question of how responsible the black block is for the red blocks staying on the table. It does so by simulating counterfactual rollouts of what would happen if the black block were removed. The CSM runs a large number of simulations, each time applying noise in several ways to capture different types of uncertainty that the observer has about the scene. After each simulation, the identities of the red blocks that have fallen off the table is recorded. The CSM assumes that the extent to which the black block is responsible for the red blocks staying on the table is linearly related to the average proportion of red blocks that would fell off the table, across all simulations. The CSM considers three sources of uncertainty that influence an observer's mental simulations about what would happen. Each source of uncertainty is modeled by introducing a small amount of random noise into the simulation. **a)** *Perceptual noise* translates all blocks horizontally by a small amount. **b)** *Intervention noise* applies impulses to blocks that are above the black block. **c)** *Dynamic noise* perturbs the normal forces applied to blocks during collisions. For each source of uncertainty, one free parameter (β_p, β_i, and β_d) determines how much noise is applied. You can play around with the CSM's parameters and see it in action here: `https://cicl-stanford.github.io/mental_jenga/docs/interface`

the blocks to the left, and some to the right. The magnitude of this perturbation is sampled from a Gaussian distribution $\mathcal{N}(0, \beta_p)$ independently for each block. Moving the blocks this way will cause some shifting of relative positioning and contact points. To ensure that the scene is still stable, we follow the method of Battaglia et al. (2013) and run the physics engine for several steps after the blocks were moved to allow the scene to settle. After allowing the scene the settle, we put all of the blocks to "sleep" (see Ullman et al., 2017). The idea is simple: if something isn't moving, the physics engine doesn't have to worry about it until something makes it move. So rather than constantly simulating the positions, velocities, and forces applied to all objects in a scene, the physics engine only tracks objects that are currently in motion. Objects that are asleep are woken up by making contact with other objects.[2]

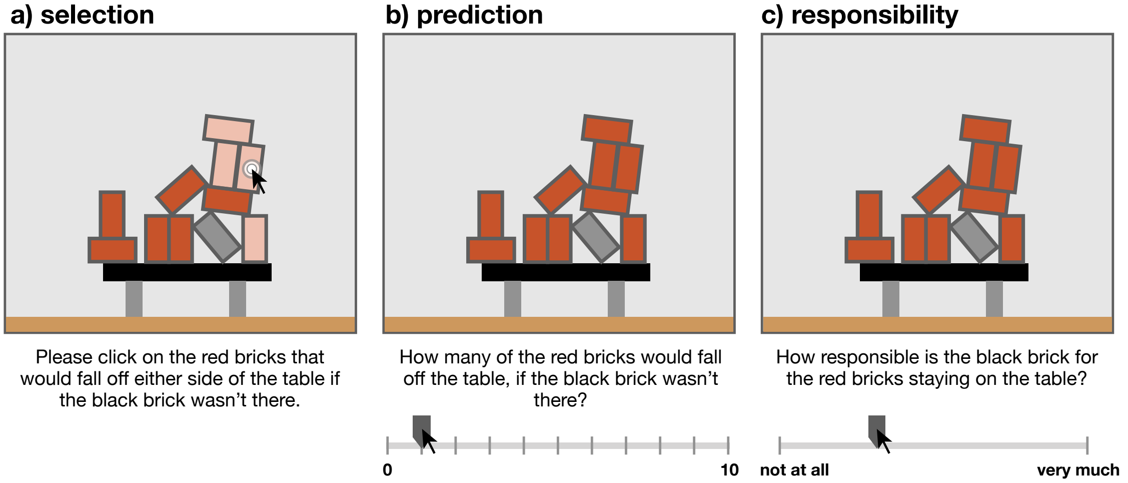
Intervention uncertainty. In addition to uncertainty about the blocks' positions, the CSM also assumes uncertainty about how the counterfactual intervention would occur. In our experiments, we ask participants to consider what would happen if the black block weren't there. Figure 2b illustrates how we implemented intervention uncertainty. First, the black block is removed from the scene simply by making it disappear. Then, a random impulse is applied to all the blocks that were above the black block. This captures the idea that, like in Jenga, the blocks *above* the intervened-on block are most directly affected by its removal. Figure 2b illustrates our criteria for whether one block is above another. In this example, block A is above the black block because 1) the two blocks contact each other, and 2) *at the contact point between the two blocks*, block A is on top of the black block. This rule is then applied recursively to find all blocks that are above the black block. So in this same example, blocks A, B and C are above the black block, and a random impulse is applied to each of them after the black block is removed.

The magnitude of the impulses applied to each block are drawn independently from $\Gamma(\beta_i, 1)$ where $\Gamma(k, \theta)$ is a Gamma distribution with shape parameter k and scale parameter θ. The angles of impulses are drawn uniformly at random from $[\frac{\pi}{4}, \frac{3\pi}{4}]$, with $\frac{\pi}{2}$ being vertically upwards. This means that impulses are usually small and directed in a roughly upwards direction, mimicking the disturbance that would be caused if the black block were manually removed from the scene (see the dotted arrows in Figure 2b).

Dynamic uncertainty. After the black block is removed, the physics engine will deterministically simulate the dynamics of how the scene will unfold. People, however are uncertain about how exactly these dynamics will play out (Allen et al., 2020; Smith & Vul, 2013). The CSM models this by adding noise to collisions, illustrated in Figure 2c. Each dotted arrow shows different samples of how the collision between the two blocks could produce different resulting trajectories. By randomly altering the ground truth normal force that results from two objects coming into contact with one another, we change how the dynamics will play out with every collision. The model perturbs the magnitude of that force: for a normal force expressed in polar coordinates as $\mathbf{F} = (F, \theta)$, we alter it so that $\mathbf{F'} = (\alpha F, \theta)$ where $\alpha \sim \mathcal{N}(1, \beta_d)$. This means that the noisy perturbation is proportional to the original force magnitude.

We will show below that the CSM that includes these different sources of uncertainty captures people's judgments to a high degree of quantitative accuracy. Of course, this doesn't mean that these are the only plausible sources of uncertainty. For example, the model doesn't capture people's uncertainty about underlying physical parameters such as the level of friction, or the bounciness of the blocks. It is possible that an alternative

[2] Setting the objects to sleep also avoids artifacts in the physics engine that may arise from continuously resolving collisions between objects. For example, initially stationary objects might start moving because there is some randomness in how the contacts between the blocks are being resolved. Setting objects to sleep prevents that from happening.



a) selection

Please click on the red bricks that would fall off either side of the table if the black brick wasn't there.

b) prediction

How many of the red bricks would fall off the table, if the black brick wasn't there?

c) responsibility

How responsible is the black brick for the red bricks staying on the table?

Figure 3. Schematic of the different experimental conditions in Experiments 1 and 2. Participants were asked to either **a)** select which blocks would fall if the black block wasn't there by clicking on the blocks (selection), **b)** judge on a slider how many blocks would fall (prediction), or **c)** judge on a slider how responsible the black block is for the red blocks staying on the table (responsibility).

noisy simulation model captures participants' judgments even better. However, our goal is not so much to determine exactly what sources of noise best capture people's predictions about what would happen. Instead, our main focus is to establish the relationship between counterfactual simulation and responsibility judgments. The model we illustrate here is just one proposal for how these counterfactual simulations may play out. We will return to the question of how other sources of uncertainty may affect people's mental simulations in the General Discussion.

From counterfactual simulations to predictions and responsibility

In our experiments, we probed participants' physical scene understanding in three different ways (see Figure 3). In the *selection condition*, participants selected which blocks would fall off the table if the black block wasn't there. In the *prediction condition*, participants indicated how many blocks would fall off the table. In the *responsibility condition*, participants judged to what extent the black block was responsible for the red blocks staying on the table.

Figure 2 illustrates how the CSM yields graded predictions about how likely the different blocks would fall off the table if the black block were removed. The model begins with an accurate encoding of the scene.[3] It then simulates the removal of the black block under different sources of uncertainty as described above. Each noisy simulation yields a

[3] We start with an accurate encoding because we assume that visual encoding is roughly accurate (subject to perceptual uncertainty), and use reasonable and constant values for properties such as density, friction, elasticity, and gravity. We also familiarize participants with the world and blocks beforehand to allow them to calibrate their assumptions. The full specification of the physical parameters can be accessed in the online materials here: `https://github.com/cicl-stanford/mental_jenga/blob/master/code/js/params.md`

different result. For example, in one simulation a particular block may fall off the table, whereas in another simulation that same block may remain on the table. The model runs many of these noisy simulations and records for each block in each simulation whether or not it fell. The model's graded prediction about whether a particular block will fall is then simply the proportion of times in which this block fell across the noisy simulations (see Figure 2, top right).

The CSM uses the proportion of times with which each block fell off the table across the noisy simulations to predict which blocks participants will select in the *selection condition*. To model participants' predictions in the *prediction condition*, the CSM uses the average number of blocks that fell across all of the simulations. Finally, to model participants' responsibility judgments in the *responsibility condition*, the CSM considers a linear mapping from the proportion of blocks that would fall off the table. So if 3 out of 4 remaining blocks were to fall off the table, the removed block would be more responsible than in a situation in which 4 out of 10 remaining blocks were to fall off.[4]

We fitted the noise parameters of the CSM (β_p, β_i, and β_d) to participants' selections and tested the full model against lesioned versions that only include a subset of the noise parameters, as well as a features model (described below) via cross-validation. The stimuli were implemented with Box2D (`https://www.npmjs.com/package/box2d`) and visualized with IvanK (`http://lib.ivank.net`). The physics simulations, including the removal of the black block and addition of different types of noise, were performed with Box2D's engine. Further details about the implementation are available online at `https://github.com/cicl-stanford/mental_jenga`.

Features model

As an alternative to the CSM, we consider a *features model* that captures participants' judgments based on visual features of the scene. Instead of simulating what would happen if the black block were removed, this model makes predictions about how likely individual red blocks are to fall by fitting a logistic regression from a collection of features to participants' responses. It then uses a linear mapping from the proportion of blocks that would fall off the table to predict participants' responsibility judgments, just like the CSM.

Table 1 shows what features the model uses to predict whether or not a red block would fall if the black block were removed. There are three categories of features: *Scene features* capture aspects of the whole scene such as how many blocks were present. *Black block features* capture aspects about the black block such as its vertical position and how many blocks are above it. *Other block features* capture aspects about the other (non-black) blocks such as their distance from either edge of the table. Table 2 shows how well each of the chosen features correlate with participants' selections in Experiments 1 and 2, and with participants' predictions in Experiment 3 about a specific (white) block.

To compare the CSM and the features model with participants' judgments across the three experiments reported below, we fit each model's parameters using cross-validation. We provide details on how the cross-validation was implemented in a separate section just before the General Discussion. When we report the experiment results below, we com-

[4] Using the proportion of blocks that would fall is consistent with Battaglia et al. (2013) who also mapped the proportion of blocks that will fall across the simulations to people's predictions about the tower's stability.

pare participants' judgments with the versions of the CSM and the features model that best accounted for participants' judgments across the cross-validation runs that incorporate participants' responses from all three experiments. By comparing the features model and the CSM, we can assess whether mental simulation is necessary for capturing people's judgments in the tasks. If the CSM outperforms the features model, as we hypothesize, then this would be consistent with the idea that people rely on mental simulations in their judgments. On the other hand, if a features model manages to perform as well as the CSM does, then this would suggest that people might arrive at their judgments without mental simulation.

Experiment 1: Investigating general stability judgments

In this experiment, participants viewed a variety of block towers like the ones shown in Figure 4. The experiment had two main goals. First, we wanted to test to what extent the CSM and the features model can capture people's beliefs about which blocks would fall off the table if the black block weren't there. The results of this comparison provide insight into the role of mental simulation in people's judgments. Second, we wanted to test the purported relationship between counterfactual predictions and judgments of responsibility. We predicted a close mapping between the counterfactual predictions of one group

Table 1
List of features used by the features model*. The 'type' column indicates whether the feature captures something about the scene, the black block, or the other blocks. The "other" blocks refer to the red blocks in Experiment 1 & 2, and to the white block in Experiment 3. In experiment 3, the "other" white block was always in the same pile as the black block.*

type	name	description
scene	`n_blocks`	total number of non-black blocks in the tower
	`avg_y`	average vertical position of the blocks in the tower
	`avg_edge_dist`	average horizontal distance of each block from the nearest table edge
	`avg_angle`	average angular deviation of each block from either a fully horizontal or vertical position
black	`black_y`	vertical position of the black block
	`black_edge_dist`	horizontal distance of the black block from the nearest table edge
	`black_angle`	angular deviation of the black block from either a fully horizontal or vertical position
	`black_above`	number of blocks above the black block
other	`other_y`	vertical position of the block
	`other_edge_dist`	horizontal distance of the block from the nearest table edge
	`other_angle`	angular deviation of the block from either a fully horizontal or vertical position
	`other_black_pile`	whether the block is in the same pile as the black block

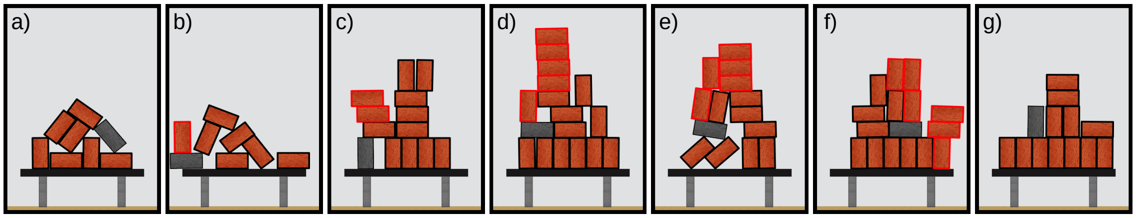


Figure 4. **Experiment 1**. Example stimuli. Red outlines indicate blocks that would fall off the table if the black block weren't there and black outlines indicate the blocks that would stay on the table. The outlines were not shown in the experiment.

of participants, and the responsibility judgments from another group of participants. The greater the proportion of blocks was predicted to fall if the black block weren't there, the more responsibility should be assigned to the black block.

Methods

All experiments reported in this paper have received approval by MIT's institutional review board (COUHES #0812003014: Learning and Reasoning with Words and Concepts).

Table 2
Correlation coefficients between individual features and participants' selection judgments for each of the three experiments. Note: *The* scene features, black block features, other block features, *and* all features *rows show how well regressions that combine these features correlate with participants' judgments. See Table 1 for a description of each feature.*

	Experiment 1	Experiment 2	Experiment 3	All
`n_blocks`	−.03	−.13	.20	−.06
`avg_edge_dist`	.12	−.05	−.07	.09
`avg_angle`	.04	.12	−.04	.13
`avg_y`	.16	−.01	.09	.08
scene features	**.23**	**.15**	**.23**	**.23**
`black_y`	−.02	−.31	−.17	−.23
`black_edge_dist`	.03	.13	−.26	.10
`black_angle`	.05	−.02	−.29	.05
`black_above`	.10	.37	.05	.28
black block features	**.12**	**.39**	**.41**	**.30**
`other_y`	.68	.39	.57	.53
`other_edge_dist`	−.08	−.27	−.21	−.13
`other_angle`	.24	.00	.18	.16
`other_black_pile`	.04	.19	-	.17
other block features	**.75**	**.56**	**.78**	**.63**
all features	**.78**	**.69**	**.84**	**.71**

Participants. 121 participants ($\text{M}_{\text{age}} = 34$, $\text{SD}_{\text{age}} = 12$, 74 male, 47 female) were recruited via Amazon Mechanical Turk using psiTurk (Gureckis et al., 2016) and randomly assigned to one of the three experimental conditions: *selection* ($N = 38$), *prediction* ($N = 42$), and *responsibility* ($N = 41$). We excluded participants from further analysis based on a catch trial which is described below. No participants failed the catch trial in the selection condition, eleven participants failed in the prediction condition (leaving $N = 31$), and six participants failed in the responsibility condition (leaving $N = 35$).

Design. Experiment 1 consisted of three conditions illustrated in Figure 3. In the *selection condition* (Figure 3a), 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." Participants were free to select any number of blocks. They could also select no blocks if they believed that none would fall. In the *prediction condition* (Figure 3b), participants were asked: "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 blocks present in the scene in steps of 1. For example, for the tower shown in Figure 4a the slider ranged from 0 to 7 whereas for Figure 4c it ranged from 0 to 12. In the *responsibility condition* (Figure 3c), participants were asked: "How responsible is the black brick for the red bricks staying on the table?" They responded on a sliding scale that ranged from "not at all" to "very much" (coded from 0 to 100 for the purpose of analysis).

Procedure. The procedure for all three conditions was largely identical. Participants first received instructions about the task. They then saw a number of warm-up animations that showed twenty blocks being dropped on the table from above. These animations were shown to familiarize participants with the relevant physical properties such as gravity, the friction between the blocks and the table, and the elasticity that influences how the block collisions play out. Participants proceed to the next stage once they had watched at least five animations. In order to go to the main experiment phase, participants had to successfully answer a comprehension check question about the task. If they answered the comprehension check question incorrectly, they were redirected to the instructions.

After the instruction phase, participants saw 42 images of different towers of blocks in randomized order (see Figure 4 for examples). The stimuli varied the number of blocks on the table (mean = 13.7, SD = 3.3, range = 7 to 20), as well as the number of red blocks that would fall off the table if the black block were removed according to ground truth (mean = 2, SD = 1.9, range = 0 to 6). Participants' tasks differed depending on the condition as described above. The experiment included a catch trial in which the black block was standing on its own (shown in Figure 4g) that we used as an exclusion criterion. Participants were excluded from the analysis if they selected that one of the red blocks would fall in that trial, if they predicted that one or more blocks would fall, or if they assigned a responsibility value greater than 15. At the end of the experiment, participants were asked to provide open-ended feedback about the task, and provided demographic information. On average, the experiment took 15.71 minutes (SD = 8.31) to complete in the selection condition, 9.86 minutes (SD = 6.49) in the prediction condition, and 8.88 minutes (SD = 8.90) in the responsibility condition.

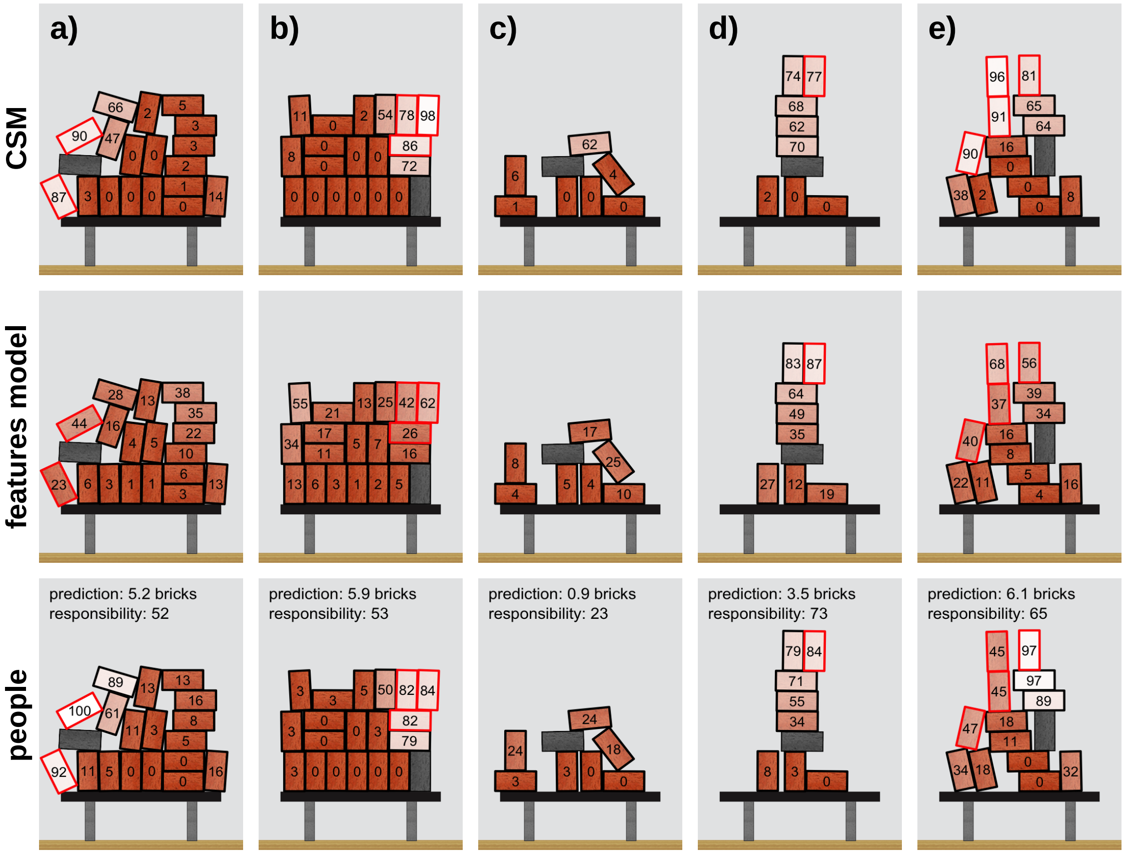


Figure 5. **Experiment 1**: Participants' selections of which blocks would fall (bottom row) together with the predictions of the *features model* (middle row) and the *counterfactual simulation model* (CSM, top row). The numbers on each block indicate the percentage of participants who thought that this block would fall off the table if the black block were removed (bottom row), or the predictions by the two different models (middle and top row). The bottom row also shows (in text) the average number of blocks that participants predicted would fall, and how responsible the black block was judged for the others to stay on the table. *Note*: The color fill gradient of the blocks maps onto 0 (red) and 100 (white). As in Figure 5, a red border indicates that a block would fall off the table according to the ground truth. The outlines were not displayed in the experiment.

Results

Figure 5 shows participants' responses for a selection of trials together with the predictions of the CSM and the features model. For each trial, the top row shows the CSM predictions, the middle row shows the features model predictions, and the bottom row shows aggregated participant responses. We will now discuss the results from each condition in turn, using these trials as illustrative examples.

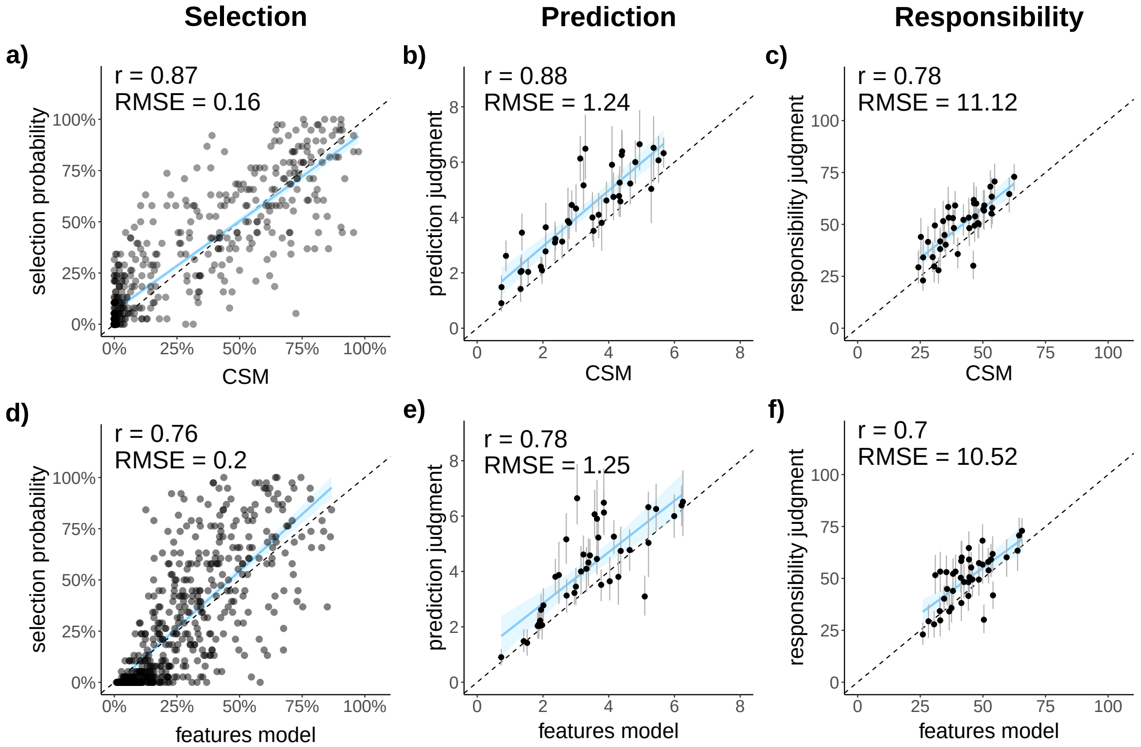


Figure 6. **Experiment 1**: Scatterplots showing the relationship between the CSM and participants' judgments at the top, and the relationship between the features model and participants' judgments at the bottom. The first column shows the results from the selection condition. Here, each data point represents the probability that one particular block was selected to fall off the table (523 blocks in total across 42 trials). The second column shows the results from the prediction condition. Here, each data point represents the average number of blocks that was predicted to fall in each trial. The third column shows the results from the responsibility condition. Here, each data point represents the average responsibility that was assigned to the black block in that trial. The blue line in each plot indicates the best-fitting regression line, and the blue ribbon shows the 95% confidence interval of the regression line. The error bars on the data points indicate 95% bootstrapped confidence intervals.

Selection condition. In the selection condition, participants were asked to click on each block that would fall off the table if the black block weren't there. The numbers on the blocks in the bottom row of Figure 5 shows the percentage of participants who selected each of the different blocks for five of the trials. For example, in the trial shown in Figure 5a, 92% of the participants selected the block on the left edge of the table, and only 16% of participants selected the block on the right side of the table. The top row in Figure 5 shows the CSM's predictions, and the middle row shows the predictions of the features model. Both models capture participants' responses in some trials, but not in others. For example, in Figures 5a and 5b, the CSM's predictions closely match participants' selections while the features model's predictions aren't as accurate. On the other hand, in Figure 5c, the

CSM assigns a high probability that the block on top would fall, whereas participants don't believe so (and this is accurately captured by the features model). In Figure 5d both the CSM and the features model closely match participants' selections. In contrast, in Figure 5e both models fail to match participants' responses. The CSM assigns a high probability that the blocks on the top left would fall whereas relatively fewer people think they would, while the features model assigns a low probability that the blocks on the top right would fall whereas relatively more people think they would.

Figures 6a and 6d show how well the CSM and the features model capture participants' selections across all of the trials. Overall, the CSM does a better job of fitting participants' selections than the features model. However, each model is somewhat biased in its predictions about blocks that people think are unlikely to fall. There are a number of blocks for which the CSM is certain that they wouldn't fall but for which participants believe have some chance of falling (see the black dots in the bottom left corner in Figure 6a extending from y = 0% to y = 25%). In contrast, the features model tends to predict that blocks *would* fall for which participants are fairly certain that they won't (see the black dots in the bottom left corner in Figure 6d extending from x = 0% to x = 25%).

To get a sense for how well participants were doing in the task, we calculated their accuracy. The overall accuracy is given by the percentage of times in which a participant correctly selected a block that falls, and didn't select a block that didn't fall. Participants' selections were 77% accurate (67% for blocks that would fall, and 79% for blocks that wouldn't fall). The CSM's accuracy was 79% (63% would fall, 83% wouldn't fall) and the features model's accuracy was 72% (47% would fall, 77% wouldn't fall).

Prediction condition. In the prediction condition, participants were asked to predict how many red blocks would fall off the table if the black block weren't there. The CSM which best accounted for participants' selections, also captures participants' judgments of how many blocks would fall (Figure 6b). The features model does not capture participants' prediction judgments as well (Figure 6e).[5]

Figure 7a shows the relationship between the average proportion of blocks that participants selected in the selection condition and the proportion of blocks predicted to fall in the prediction condition. Overall, the two ways of probing participants yielded very similar results. However, participants in the prediction condition tended to predict that a larger proportion would fall than participants in the selection condition selected (as indicated by the fact that the regression line in Figure 7a is slightly above the the diagonal).

Responsibility condition. In the responsibility condition, participants were asked to judge the extent to which the black block was responsible for the red blocks staying on the table. To account for the fact that different scenes have a different number of blocks, we used the *proportion* of blocks predicted to fall as a predictor for people's responsibility judgments. We calculate the proportion simply by dividing the number of blocks predicted to fall by the total number of red blocks in the scene. Because it's the proportion rather than the absolute number that matters, this means that a block's responsibility is greater in a scene where, for example, three out of four blocks are predicted to fall compared to a scene in which four out of ten blocks are predicted to fall.

[5] Instead of reporting frequentist statistics to compare the models here, in the 'Parameter fitting and model comparison' section before the General Discussion, we report the results of a cross-validation that compares how well the different models do across the three experiments in this paper.

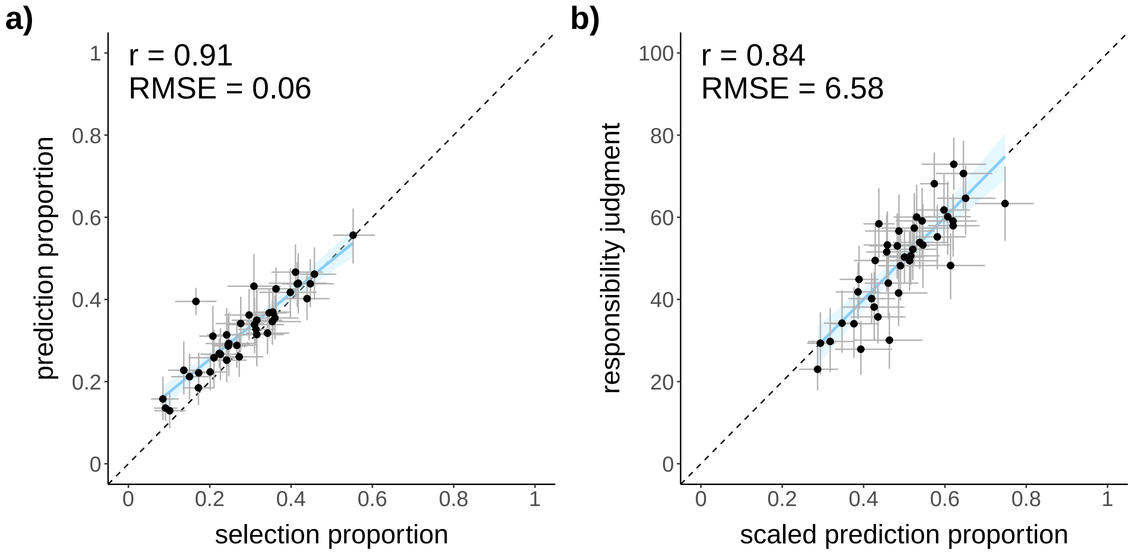


Figure 7. **Experiment 1**: Comparison between participants' responses in the three conditions. **a)** The proportion of blocks participants selected in the selection condition (x-axis) is closely related to the proportion of blocks predicted to fall in the prediction condition (y-axis). **b)** The (scaled) proportion of blocks predicted to fall in the prediction condition (x-axis) is closely related to participants' responsibility judgments (y-axis). The scaling here is done via a simple regression that maps from the selection proportion (which ranges between 0 and 1) to participants' responsibility judgments (which ranges between 0 and 100). The greater the proportion of blocks is predicted to fall, the more responsible the black block is judged to be. *Note:* The error bars indicate bootstrapped 95% confidence intervals, and the blue ribbons show the 95% confidence interval of the regression lines.

We fit a linear regression from the proportion of blocks predicted to fall to participants' responsibility judgments. Figure 7b shows that the counterfactual predictions from one group of participants are closely related to the responsibility judgments from another group of participants (r = .84). This result is consistent with the idea that when evaluating how responsible the black block is, participants consider what proportion of other blocks would fall if it were removed.

We apply the same linear transformation that maps from participants' predictions to responsibility judgments for both the CSM and the features model. Figure 6c shows how well the CSM captures participants' responsibility judgments, and Figure 6f shows the same plot for the features model. Both models do a similarly good job in capturing participants' judgments.

Discussion

The results of Experiment 1 reveal a close mapping between counterfactual predictions and responsibility judgments. The greater the proportion of red blocks were predicted to fall, the more responsible the black block was judged to be (Figure 7b).

We tested participants' ability to simulate what would happen if the black block were removed in two different conditions (Figure 3a and 3b). In the selection condition, participants clicked on each red block that they believed would fall if the black block weren't there. In the prediction condition, participants indicated how many red blocks would fall if the black block was weren't there. The number of blocks that participants selected and the number of blocks they predicted to fall were highly correlated (Figure 7b).

Compared with the ground truth, participants in the prediction condition were less accurate (RMSE = 2.64) than participants in the selection condition (RMSE = 2.29). Having to decide for each block whether or not it's going to fall is likely to lead to a more careful consideration of what would happen than having to merely move a slider to estimate how many blocks would fall. Participants in the selection condition also spent considerably more time to complete the experiment compared with participants in the prediction condition.

We compared participants' responses to the predictions from both the CSM and the features model. The CSM assumes that participants use their intuitive understanding of physics to simulate what would happen to the red blocks if the black block weren't there (Figure 2). The features model assumes a direct mapping from visual features to participants' responses. We found that across the three conditions, the CSM was overall a better fit to experimental data, suggesting that counterfactual simulations may be critical for explaining participants' judgments.

Even though the CSM performed better overall, it's still possible that participants may have mostly relied on perceptual features in their judgments. For example, the average y-position of the blocks in the scene alone correlated highly with participants' responsibility judgments (r = .71, see Table A1 in the Appendix). To provide stronger evidence for the role of mental simulation in people's judgments about physical stability and to investigate peoples' judgments when scene features are controlled for, we constructed the block towers differently in Experiment 2.

Experiment 2: Controlling for scene features

In Experiment 1, we tested participants' judgments on a wide array of randomly generated towers, and the features model captured participants' responsibility judgments quite well by considering global scene features, such as the average y-position of the blocks in the scene. In Experiment 2, we used a more tightly controlled stimulus set. We wanted to make sure that global scene features would not be highly correlated with the number of blocks that would fall. We achieved this by first constructing a set of six different tower configurations by hand. For each configuration, we then chose seven positions for the black block such that removing it would result in different numbers of blocks falling off the table in the ground truth setting.

Figure 8 shows a subset of the stimuli that we used in this experiment. Relying on scene features to predict how many blocks would fall is now insufficient: in fact, the scene features for any of the instances of a given tower configuration are identical (e.g. within any of the Towers in Figure 8) as only the position of the black block varies. Furthermore, while in Experiment 1 the blocks in each stimulus tended to form a single “stack” (see Figure 4), in Experiment 2 we created some tower configurations with disjointed sets of blocks. For example, Towers I, II, and IV in Figure 8 feature two sets of blocks that are disconnected from one another. In the scene shown in Figure 8a, for instance, it is clear that the removal

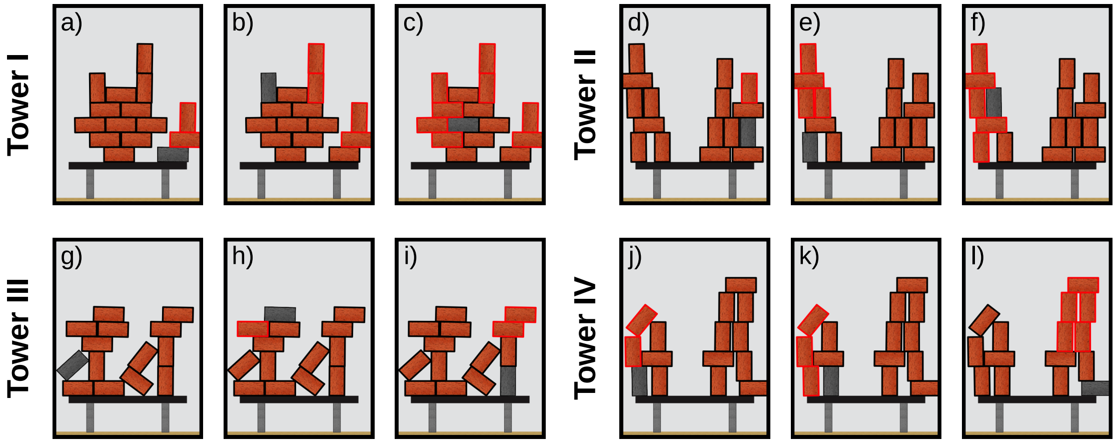


Figure 8. **Experiment 2**. Example stimuli. We created six different tower configurations (four shown), each of which was repeated seven times for different positions of the black block (three shown). As before, red outlines indicate which blocks would fall off the table if the black block weren't there. Outlines were not visible to participants in the experiment.

of the black block should only affect the two red blocks above it. Overall, this new set of stimuli provides a stronger test for the potential role of mental simulation in participants' responsibility judgments.

Methods

Participants. 129 participants ($M_{age} = 36$, $SD_{age} = 11.3$, 70 male, 59 female) were recruited via Amazon Mechanical Turk with $N = 44$ in the selection condition, $N = 42$ in the prediction 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 4g. One participant was excluded in the selection condition (leaving $N = 43$), two were excluded in the prediction condition (leaving $N = 40$), and three were excluded in the responsibility condition (leaving $N = 40$).

Design & Procedure. The design, procedure, and questions were identical to those of Experiment 1. The main difference was the set of tower stimuli that we used this time (compare Figure 8 with Figure 4). We reduced the table friction in the settings of the physics engine so that it was possible for blocks to slide off the table. Participants saw 43 trials in randomized order where one trial served as a catch trial (see Figure 4g). On average, the experiment took 13 minutes (SD = 6.87) to complete in the selection condition, 11.6 minutes (SD = 5.24) in the prediction condition, and 7.86 minutes (SD = 3.48) in the responsibility condition.

Results

Figure 9 shows participant responses and model predictions for a selection of stimuli. As before, the top row shows the CSM predictions, the middle row shows the features model predictions, and the bottom row shows aggregated participant responses for each block in the scene.

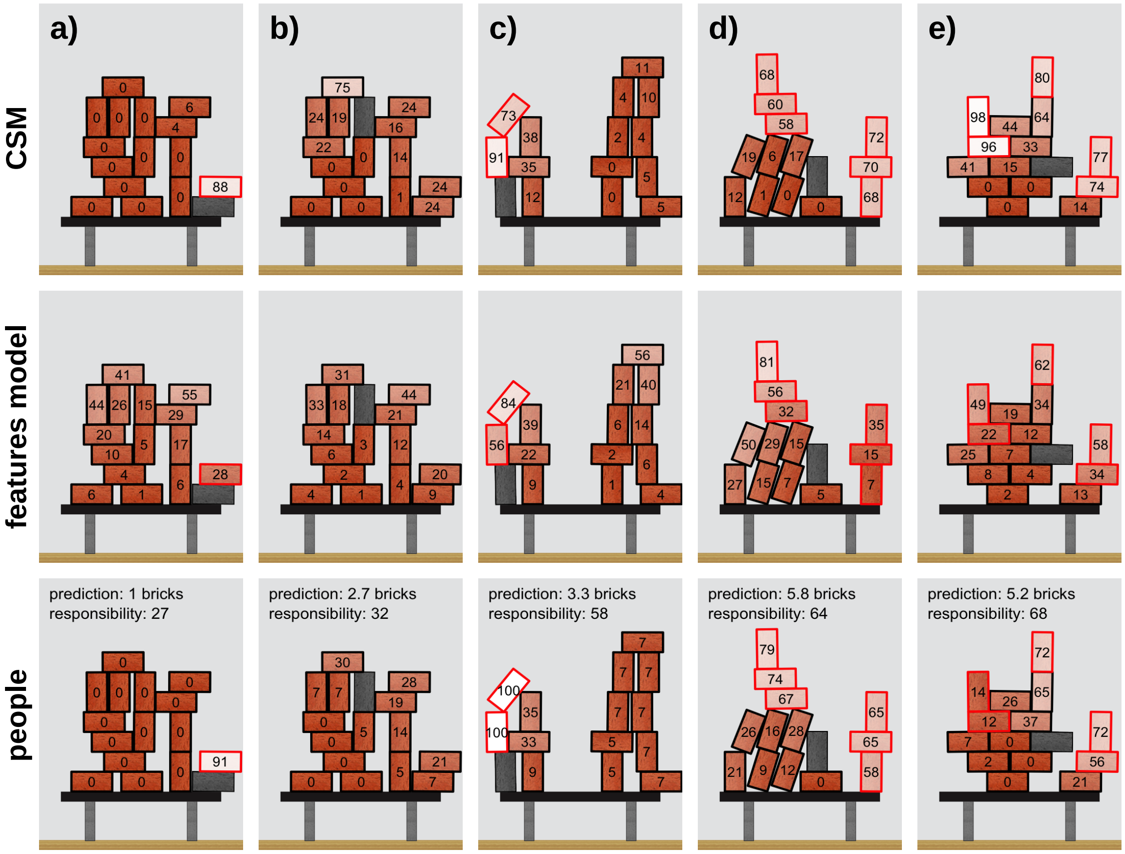


Figure 9. **Experiment 2:** Model predictions and participants' judgments for a selection of stimuli. *Note*: The number on each block indicates the percentage of participants who thought that this block would fall if the black block weren't there, and the predicted percentages for the models. The color fill gradient of the blocks maps onto 0 (red) and 100 (white). A red border indicates that a block would fall off the table, and a black border indicates that a block would remain on the table. The outlines were not displayed in the experiment.

Selection condition. As in Figure 5, the selection of stimuli in Figure 9 includes examples where the models accurately capture participants' selections as well as examples where the models don't perform as well. For example, in Figure 9a participants realized that none of the blocks apart from the one immediately above the black block would fall off the table if the black block weren't there. The CSM captures participants' selections well here, whereas the features model doesn't. The features model assigns a relatively high probability that blocks in the middle of the scene would fall, and a low probability that the block just above the black block would fall. More generally, the features model tends to overestimate that blocks would fall which participants don't select, and underestimate the ones that participants do select.

Figure 9b shows an example where the block tower configuration is the same as in

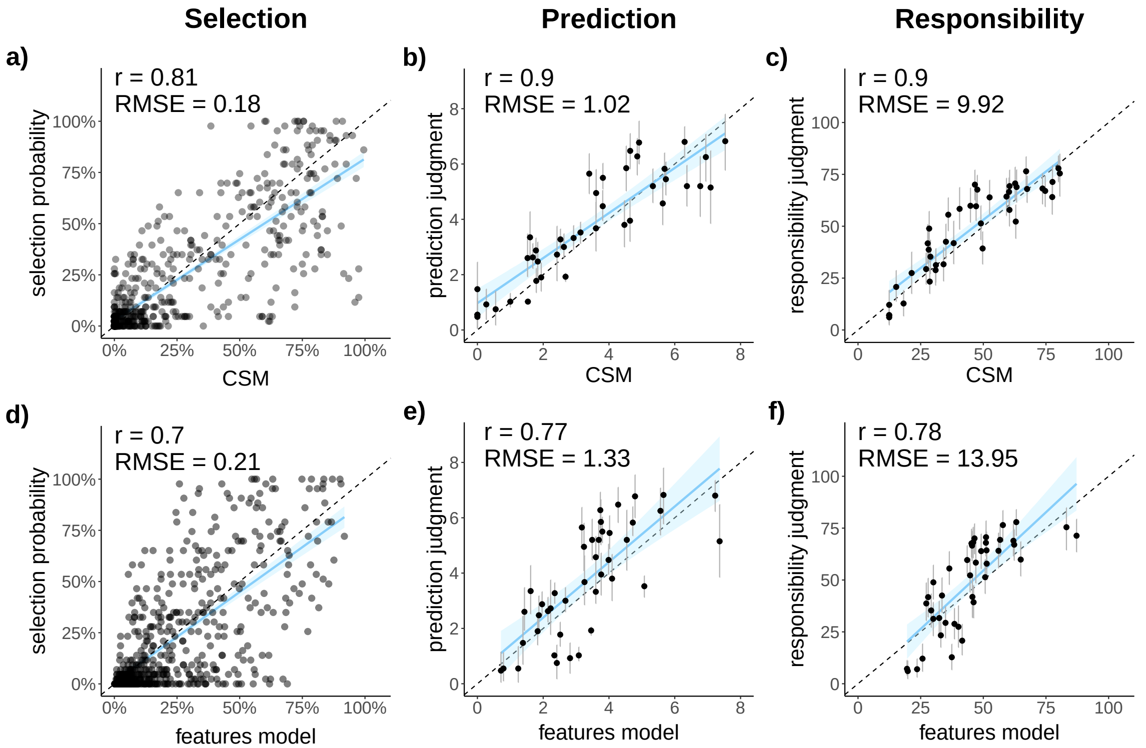


Figure 10. **Experiment 2**: Scatterplots showing the probability with which a block was selected on the y-axis and the predictions of the CSM (top) or the features model (bottom) on the x-axis, for each of the three conditions. Each point in a) and d) represents one block in one of the trials (630 blocks in total). Each point in the remaining panels represents one trial (42 trials in total). The blue line in each plot indicates the best-fitting regression line, and the blue ribbon shows the 95% confidence interval of the regression line. The error bars on the data points indicate 95% bootstrapped confidence intervals.

Figure 9a, but the position of the black block is different. Naturally, the position of the black block makes a big difference to participants' selections, and the CSM correctly captures this. However, the features model's predictions about which blocks would fall are very similar in Figure 9a and Figure 9b. Even though the black block features are changed between these scenes, the global features, and the features about the red blocks are identical, leading the model to make similar predictions in both cases (see Table 1).

Figure 9c shows an example where two sets of blocks are disconnected from one another. Here participants didn't think that the blocks on the right side would fall and the CSM captures this correctly. In contrast, the features model assigns a high probability that these blocks would fall. Figure 9d shows an example for which both the CSM and the features model capture people's selections well. In Figure 9e, the features model does a better job than the CSM. The CSM assigns a high probability that the two blocks on the top left would fall (which is what would happen according to ground truth). However, participants didn't think that these blocks would fall.

Figures 10a and d show how well the CSM and the features model capture partici-

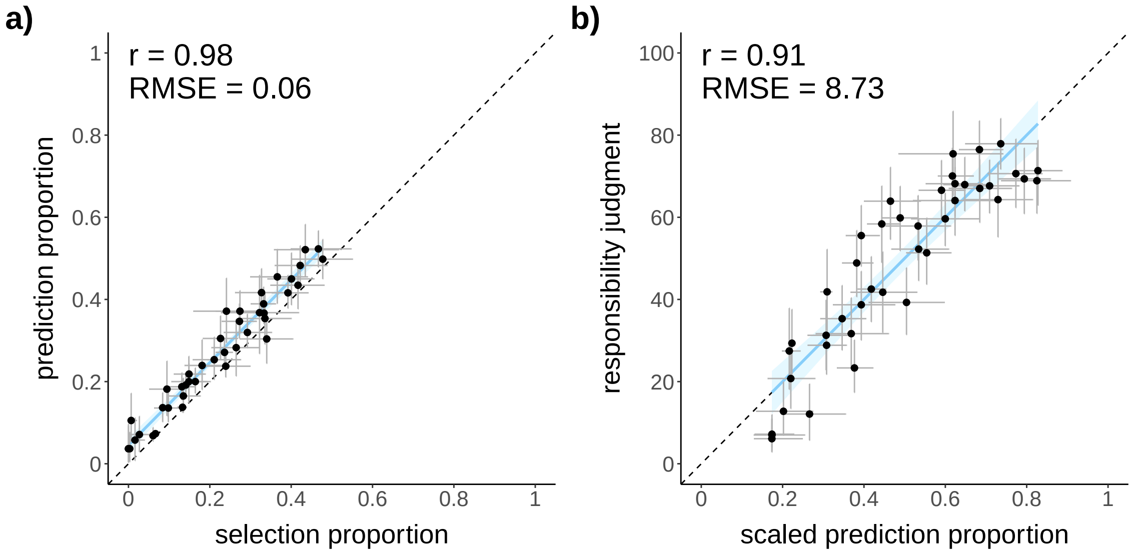


Figure 11. **Experiment 2**: Comparison between participants' responses in the three conditions. **a)** The proportion of blocks participants selected in the selection condition was compared with the proportion of blocks they judged to fall in the prediction condition. We used the proportion because different stimuli consisted of different number of blocks. **b)** The proportion of blocks predicted to fall in the prediction condition was then compared with participant's judgments in the responsibility condition. *Note:* the error bars indicate bootstrapped 95% confidence intervals. The blue ribbon shows the 95% confidence interval for the regression line.

pants' selections across all of the trials. Like in Experiment 1, the CSM performs better than the features model as evidenced by a higher correlation and a smaller error. Similarly, the features model tends to predict that blocks would fall for which people are certain that they wouldn't (as indicated by the many black points along the x-axis in Figure 10d). Both models tend to underestimate larger selection probabilities (the regression line is below the diagonal in Figure 10a and 10d). Compared to ground truth, participants' selections were 82% accurate (60% for blocks that would fall, and 88% for blocks that wouldn't fall). The CSM's accuracy was also 82% (62% would fall, 86% wouldn't fall), and the features model's accuracy was 73% (42% would fall, 82% wouldn't fall).

Prediction condition. The relationship between model predictions and participants' predictions about how many of the red blocks would fall if the black block weren't there are shown in Figures 10b and 10d. The CSM again does a better job of capturing participants' predictions. Figure 11a shows the relationship between the average proportion of blocks that participants selected in the selection condition and the proportion of blocks predicted to fall, just like in Figure 7a. Again, there was a very tight relationship between selections and predictions, and participants predicted that more blocks would fall on average than they selected (the regression line is above the diagonal).

Responsibility condition. Figures 10c and 10f show how well the CSM and the features model account for participants' responsibility judgments. Yet again, the CSM does a better job in capturing participants' judgments than the features model. Table 1 shows the correlations between different features with participants' responsibility judgments. As expected, global scene features did not correlate well with participants' responsibility judgments because these features are insensitive to the black block's position. This time, a good predictor of participants' responsibility judgments was the y-position of the black block. The lower the black block was located in the scene, the more responsible it was judged to be for the stability of the other blocks.

Figure 11b shows the relationship between participants' predictions and responsibility judgments. The responsibility judgments from one group of participants were well accounted for by the proportion of blocks that another group of participants predicted would fall if the black block weren't there. The greater the proportion of blocks that were predicted to fall, the more responsible the black block was judged.

Discussion

The results of Experiment 2 replicate and extend what we found in Experiment 1. Again, participants' predictions about what would happen if the black block weren't there were highly correlated with judgments about how responsible that block was for the others staying on the table.

We constructed the stimuli in Experiment 2 differently from how we did in Experiment 1. This time, we included sets of towers and manipulated within each set where the black block was positioned, while keeping everything else constant (see Figure 8). Participants' judgments in Experiment 1 were highly correlated with the average y-position of the blocks in the scene. This new way of designing the stimuli made it such that the average y-position of the red blocks was no longer a good cue because it doesn't take into account where the black block is positioned.

The scenes in Experiment 2 were also different in that they featured block towers with disconnected sets of blocks. These scenes help tease apart to what extent people's judgments are sensitive to global scene features versus the more local consequences that removing the block would have. The results showed that the CSM provided a good account of participants' judgments across all three experimental conditions, and that it outperformed the features model in each condition.

Experiment 3: Investigating the relationship between block pairs

The results of Experiments 1 and 2 showed that the CSM accurately captures participants' judgments about how responsible one block was for the tower's overall stability. The CSM also naturally makes predictions about the relationship between pairs of individual blocks, by querying what would happen to *just one* block if another were removed. The features model, in contrast, needs to be reconfigured for this novel task. In Experiment 3 we asked participants to judge how responsible one block was for another block's staying on the table. Figure 12 shows a selection of trials: each scene contained one black block, one white block, and a varying number of red blocks. In the prediction condition, participants were asked to judge how likely the white block would be to fall off the table if the black block weren't there. In the responsibility condition, participants judged to what extent the black block was responsible for the white block staying on the table.

This new task is similar to the way in which Gerstenberg, Goodman, et al. (2021) probed causal judgments (see also Gerstenberg et al., 2017). In their studies, participants were asked whether a candidate cause (e.g. ball A) caused another ball (e.g. ball B) to go through a gate, or prevented it from going through. In our case here, the question is whether the black block prevents the white block from falling off the table. Gerstenberg, Goodman, et al. (2021) asked one group of participants to make counterfactual judgments (e.g. "Would ball B have missed the gate if ball A had been removed?") and another group to make causal judgments (e.g. "Did ball A cause ball B to go through the gate?"). The results showed a very close quantitative correspondence between the counterfactual judgments of one group, and the causal judgments of another. The more certain participants were that the counterfactual outcome would have been qualitatively different from what actually happened, the more they judged that the candidate caused the outcome. Correspondingly, in our task, we expect that there will be a close mapping between the counterfactual predictions and responsibility judgments. The more certain participants are that the white block would fall if the black black weren't there, the more responsible the black block should be judged for the white block's staying on the table.

Methods

Participants. 81 participants ($\mathrm{M}_{\mathrm{age}}$ = 37.2, $\mathrm{SD}_{\mathrm{age}}$ = 12.1, 49 male, 32 female) were recruited via Amazon Mechanical Turk with $N = 41$ in the prediction condition and $N = 40$ in the responsibility condition. We used an exclusion trial in which the removal of the black block clearly had no effect on the white block, similar to the trial in Figure 4g. 3 participants were excluded in the prediction condition (leaving $N = 38$), and 3 participants were excluded in the responsibility condition (leaving $N = 37$).

Design & Procedure. The experiment instructions were largely identical to those of Experiments 1 and 2. Because we only asked participants about two particular blocks in each scene, this experiment did not include a selection condition. In the *prediction condition*, participants were asked "Would the white brick fall off the table if the black brick wasn't there?" Participants provided their answer on a sliding scale ranging from "definitely not" (0) to "definitely yes" (100). In the *responsibility condition*, participants were asked "To what extent is the black brick responsible for the white brick staying on the table?" Participants provided their answer on a sliding scale ranging from "not at all" (0) to "very much" (100).

Participants saw 42 separate scenes which had been generated in the same way as in Experiment 1 (see Figure 12 for a selection of scenes). We selected scenes such that the CSM's predictions of whether the white block would fall varied across the whole range from being certain that it wouldn't fall to being certain that it would. The number of blocks on the table ranged between 12 and 19. The number of blocks that would fall if the black block was removed according to the ground truth varied from 0 to 9. On average, participants took 7.03 (SD = 5.04) minutes in the prediction condition and 6.73 (SD = 7.99) minutes in the responsibility condition.

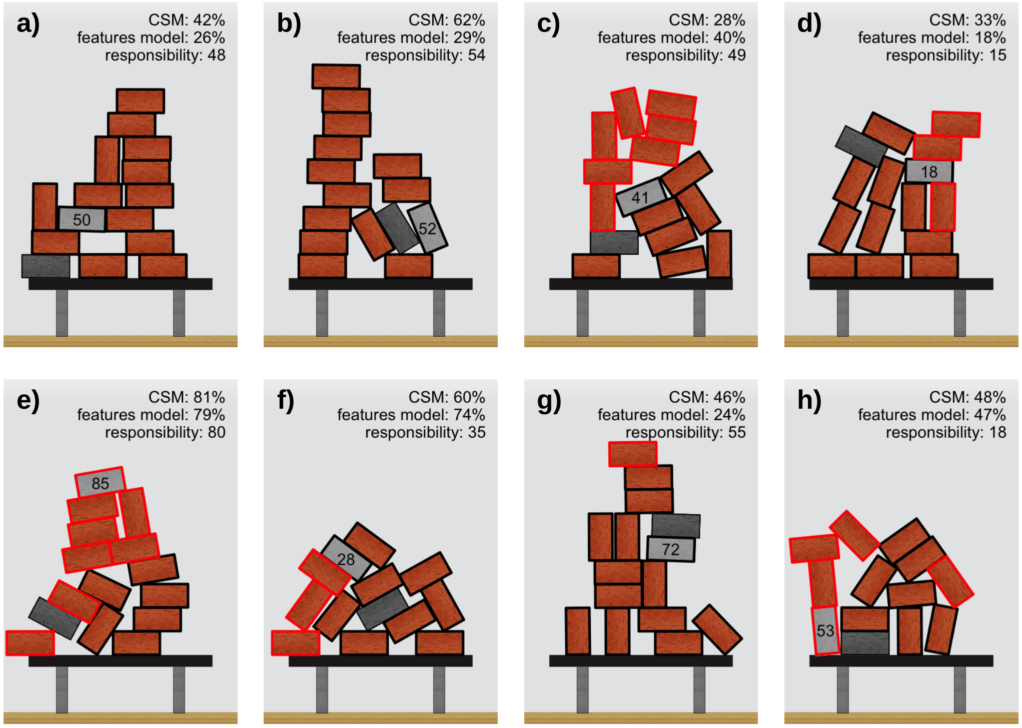


Figure 12. **Experiment 3:** Example stimuli, along with experimental data and model predictions. In the *prediction condition*, participants judged how likely the white block would be to fall if the black block weren't there. In the *responsibility condition*, participants judged how responsible the black block was for the white block staying on the table. *Note*: The number on the white block indicates participants' mean prediction judgment for this scene. The text at the top of each trial shows the CSM and the features model prediction of how likely the white black would fall, as well as participants' mean responsibility judgment. The black and red outlines indicate whether each block would stay or fall off the table if the black block weren't there; outlines were not present in the experiment.

Results

Figure 12 shows participants' responses and model predictions for a subset of the trials. The number on the white block shows participants' average *prediction* judgments. The higher the number the more likely they believed on average that this block would fall off the table if the black block were removed. The text at the top of each figure shows the predictions of the CSM and the features model, as well as participants' *responsibility* judgments. We will discuss the results of the prediction condition and the responsibility condition in turn.

Prediction condition. We asked participants how likely the white block would be to fall off the table if the black block weren't there. In Figure 12a, participants' average prediction for how likely the white block would fall off the table was 50%, which is matched quite well by the CSM (42%) but less so by the features model (26%). The same is true for Figure 12b, where the features model's emphasis on the y-position of the black and red

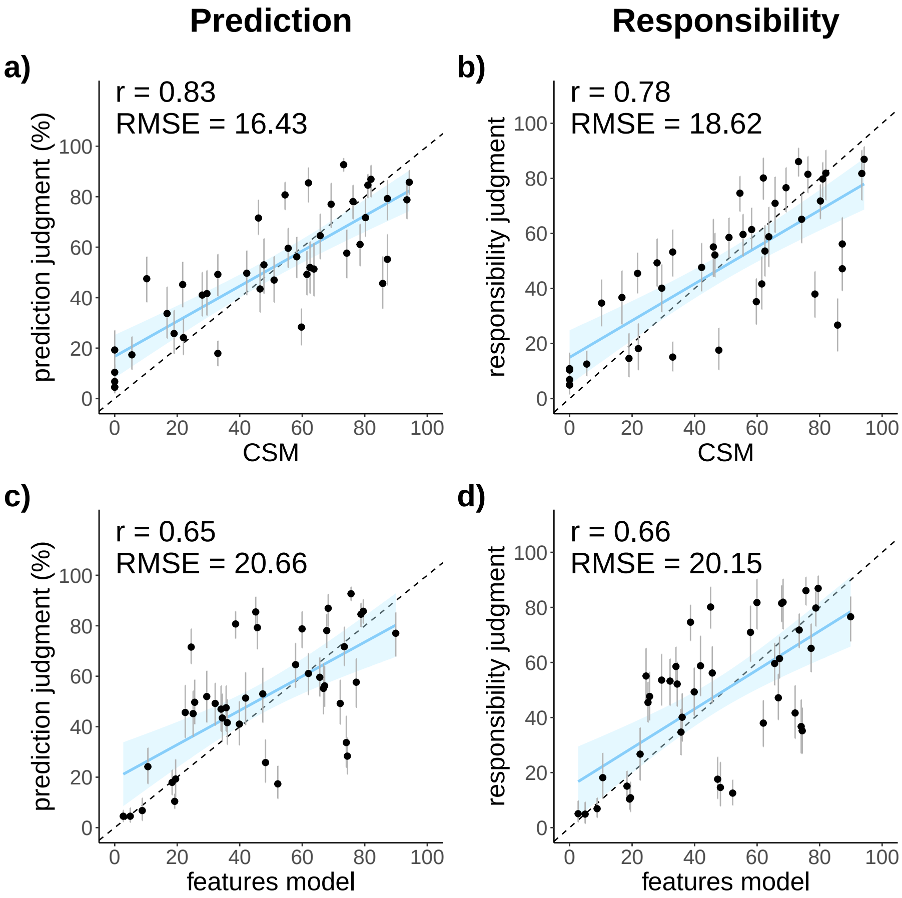


Figure 13. **Experiment 3**: Scatterplots showing the performance of the CSM and the features model on both conditions. In the prediction condition (a and c), participants were asked how likely the white block was to fall, if the black block weren't there. In the responsibility condition (b and d), participants were asked how responsible the black block was for the white block staying on the table.

block leads it to underpredict participants' judgments.

However, the features model outperforms the CSM in other cases. In Figure 12c, the white block is "shielded" from falling by the red blocks around it, but participants believed that it had a relatively high chance of falling. In Figure 12d the CSM overestimates the probability that the white block would fall compared to people's judgments. In some of the cases, both the CSM and the features model capture participants' judgments (e.g. Figure 12e). And sometimes, neither model fits participants' judgments (Figure 12f, 12g). Figure 12g is particularly interesting because the white block is directly *under* the black block. Both models perform poorly in this case – they underestimate how likely people judge that it would fall.

Figures 13a and 13c compare the predictions of the CSM and features model with participants' judgments across all 42 trials. The CSM does a better job than the features model at capturing participants' predictions. Notice that participants' judgments are less

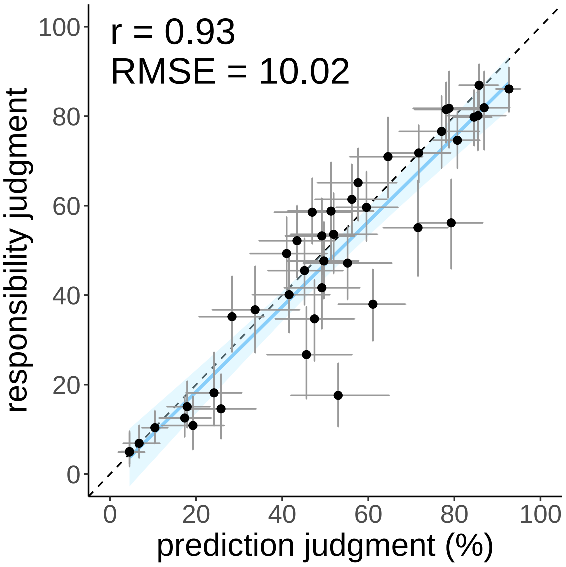


Figure 14. **Experiment 3:** Relationship between participants' predictions of how likely the white block would be to fall if the black block were removed (x-axis) and the extent to which the black block was judged to be responsible for the white block staying on the table (y-axis). Each point shows the averaged judgments for one trial, and the error bars are 95% bootstrapped confidence intervals.

extreme than what the models predict (as indicated by the regression line being off the diagonal). To get a sense for how accurate participants and the model were, we computed the average probability with which participants (or the models) said that a block would fall when it did, and that it wouldn't fall when it didn't. Participants' prediction responses were 61% accurate. The CSM and features model were 63% and 62% accurate, respectively.

Responsibility condition. Figure 14 shows the relationship between participants' judgments in the prediction condition and in the responsibility condition. The results show that there is a very close relationship between participants' predictions about whether the white block would fall if the black block weren't there, and the extent to which the black block was judged to be responsible for the white block staying on the table. The more likely participants judged that the white block would fall if the black block weren't there, the more responsible the black block was judged.

Figures 13b and 13d show how well the CSM and features model capture participants' responsibility judgments. Again, the CSM does a better job than the features model. Both models, however, fail to capture some of the variance in participants' responses.

Discussion

While Experiments 1 and 2 looked into how people judge the extent to which a candidate object is responsible for the overall stability of the tower, Experiment 3 focused on the relationship between individual blocks. We asked one group of participants to predict whether a target block (the white block) would fall if the block block weren't there, and another group of participants how responsible the black block was for the white block staying on the table. We found that participants' counterfactual predictions and responsibility judgments were closely related (see Figure 14). The more likely participants thought that the white block would fall the more responsible the black block was judged to be. There were a few cases for which prediction and responsibility judgments differed. For example, in the trial shown in Figure 12h, participants predicted that the white block would be 53% likely to fall off the table, but assigned relatively little responsibility (18) to the black block. One possibility for why the two types of judgments come apart here is that when people judge responsibility, they not only care about the chances that the other block would fall, but also about the causal chain of events by which the outcome would come about. So when several other blocks are part of the chain of events that lead from the removal of the black block to the white block falling off the table, there is a certain degree of diffusion of responsibility (see Chockler & Halpern, 2004; Gerstenberg & Lagnado, 2010; Lagnado, Gerstenberg, & Zultan, 2013; Langenhoff, Wiegmann, Halpern, Tenenbaum, & Gerstenberg, 2021; Zultan, Gerstenberg, & Lagnado, 2012, for how the causal structure affects diffusion of responsibility between agents).

In Experiments 1 and 2, the responsibility question was somewhat ambiguous. We had asked participants to judge to what extent the black block was responsible for the red blocks staying on the table. We found that participants' responsibility judgments correlated highly with the proportion of blocks that would have fallen off the table. It's possible though that different participants interpreted the question differently, such that some believed that it was the absolute number that mattered. In Experiment 3, participants had to judge how responsible one block was for another one, thereby removing this ambiguity.

Experiment 3 also connects more closely with prior work on causal judgment (see Gerstenberg, Goodman, et al., 2021). In work on causal judgment, researchers usually ask to what extent one candidate caused a particular event to happen. For example, the question might be whether a billiard ball A caused another billiard ball B to go through a gate. In this case, participants consider what would have happened if ball A hadn't been present in the scene (Gerstenberg et al., 2017). The more certain they are that the outcome would have been different in that case, the more they judge that ball A caused the outcome. In a similar way, we ask here whether one candidate, the black block, is responsible for the white block staying on the table. The results show that participants' responsibility judgments are consistent with the idea that they are mentally simulating what would happen if the black block weren't there.

The CSM again provided a good account of participants' predictions and responsibility judgments (see Figure 13). This further supports the idea that counterfactual simulation and responsibility for stability are intimately related. The features model which tries to predict participants' judgments without relying on physical simulations doesn't fare as well. In Experiment 1, scene features, such as the average distance of block from the edge, were a good predictor of participants' responsibility judgments. In Experiment 2, features associated with the black block, such as how many red blocks were above it, were a good predictor. This time, in Experiment 3, it was features associated with the white block that correlated highly with participants' responsibility judgments. So overall, while there are features in each experiment that are associated with participants' judgments, what features these are changes between experiments. In contrast, the CSM provides a unified account of participants' judgments across all experiments.

Parameter fitting and model comparison

The CSM and the features model have a number of free parameters that need to be fitted to the data. We fit the model parameters to one large dataset that combined participants' selections from Experiment 1 and 2, and their predictions from Experiment 3. The selection data provides a strong test for the models as they need to predict for each block whether participants think that it will fall or stay.

Parameter fitting

The CSM has up to three free parameters, one each for the perceptual noise, intervention noise, and dynamic noise. For any given set of the model parameters $(\beta_p, \beta_i, \beta_d)$, the CSM predicts how likely each of the red blocks would fall off the table (see Figure 2 top right). To obtain numerical predictions for each block, we ran 200 simulations for each parameter setting. We fit the CSM's parameters by maximizing the likelihood of the data, using a grid search over a wide range of possible noise parameter values. To find the best-fitting parameters for the features model, we performed a logistic regression. The features model has thirteen free parameters: one for each of the features, plus one for the intercept.

Model comparison

The full CSM includes three sources of uncertainty that affect people's predictions about what would happen. To evaluate whether all three of these components are required to accurately account for people's judgments, we compared the full model with simpler models that only consider one, or two sources of uncertainty. For example, one such model "turns off" intervention uncertainty by setting the intervention noise parameter β_i to 0.

Table 3

Cross-validation results for different versions of the counterfactual simulation model (CSM) as well as the features model. The r column shows the correlation between model predictions and participants' responses. The RMSE column shows the root mean squared error between model prediction and participants responses. The ΔRMSE column shows the difference in RMSE between the full CSM and the other models. Note*: Each column shows the median and the 10% and 90% quantiles across the 200 cross-validation runs of the respective measure.*

model	r	RMSE	ΔRMSE
CSM (p, i, d)	.84 [.82, .85]	2.95 [2.71, 3.18]	-
CSM (p, i)	.81 [.79, .82]	3.45 [3.14, 3.79]	0.50 [0.30, 0.68]
CSM (p, d)	.76 [.74, .78]	4.26 [3.94, 4.55]	1.31 [0.99, 1.61]
CSM (i, d)	.81 [.79, .83]	3.55 [3.22, 3.88]	0.60 [0.32, 0.88]
CSM (p)	.75 [.73, .77]	4.34 [4.01, 4.64]	1.39 [1.08, 1.68]
CSM (i)	.80 [.77, .81]	3.75 [3.44, 4.13]	0.80 [0.53, 1.08]
CSM (d)	.68 [.66, .71]	6.84 [6.29, 7.33]	3.89 [3.36, 4.42]
features	.72 [.70, .74]	4.34 [4.04, 4.63]	1.39 [1.03, 1.70]

p = perceptual noise, i = intervention noise, d = dynamic noise.

There are six such models (three models with two sources of noise, and three models with one source of noise).

To evaluate how well the different versions of the CSM and the features model account for participants' responses, we performed split-half cross-validation on the combined data from all three experiments. We randomly selected half of all of the blocks in all three experiments as a training set and found the best-fitting parameters for each model on the training set using the parameter fitting procedure described above. We then evaluated the model's performance on the held-out test set comprised of the other half of the blocks. For each model, we performed 200 split-half cross-validation runs. Table 3 shows the results of this analysis.

The full CSM model outperforms all of the lesioned models: it correlates higher with participants' responses in the held-out test sets, and has lower error. The cross-validation results also give a sense of how much the different sources of uncertainty affect the model's performance. For example, models that don't include intervention noise generally fare worse than those that don't include perceptual noise. The full CSM also outperforms the features model. The features model achieves a high correlation with participants' judgments when fitted separately to the different experiments (see Table 2). However, there is no single parameter setting that works well across all three of the experiments. Which features matter most differs between the experiments.

General Discussion

How do people judge whether one object supports another? In this paper, we developed the counterfactual simulation model (CSM) of physical support. The CSM predicts that people judge physical support by mentally simulating what would happen if the object of interest were removed. Similar to how people spontaneously consider counterfactuals when judging causation (Gerstenberg et al., 2017), the CSM assumes that people play "Jenga in their mind" when judging responsibility for physical support. The more certain people are that the object(s) of interest would have fallen if a target object had been removed, the more responsible that target object is for their stability.

We tested the CSM across three experiments in which participants made judgments about towers of blocks stacked on a table. In Experiments 1 and 2, the CSM accurately captured participants' *selections* of which other blocks would fall off the table if the black block weren't there, their *predictions* of how many of the blocks would fall, as well as their judgments of how *responsible* the black block was for the other blocks staying on the table. In Experiment 3, the CSM captured participants' graded beliefs about whether one particular block of interest would fall of the table if the black block weren't there. All three experiments showed how the counterfactual predictions of one group of participants closely matched the responsibility judgments of another group.

We contrasted the CSM with a features model that predicts participants' judgments via a direct mapping from visual features. For example, the features model predicts that more blocks will fall when the tower is taller. The features model wasn't able to capture participants' selections, predictions, and responsibility judgments as well as the CSM did. Which individual features best correlated with participants' judgments varied across the different experiments. In contrast, the CSM provides a unified account of participants' judgments across a wide variety of situations and tasks.

A unified account of causal judgments across different types of causation

The CSM explains people's causal judgments as arising from a comparison between the actual situation, and a counterfactual situation in which the candidate cause was imagined to have been different. The CSM has been shown to provide an accurate model of how people make causal judgments about physical events (Beller, Bennett, & Gerstenberg, 2020; Gerstenberg, 2022; Gerstenberg, Goodman, et al., 2021; Gerstenberg et al., 2017). In this standard kind of "event causation", one candidate cause event brings about an effect event of interest, such as when the rock hitting the window causes the window to shatter. Most philosophical theories of causality take the causal relata to be events (Paul & Hall, 2013; Schaffer, 2016).

Treating events as the units of a causal relationship, however, makes it difficult to handle omissions as causes. When our plants die while we were away because our neighbor forgot to water them (even though they had promised to do so), there is no event that we could attribute the outcome to (Beebee, 2004; Henne, Pinillos, & De Brigard, 2017; Livengood & Machery, 2007; McGrath, 2005). Gerstenberg and Stephan (2021) have shown that the CSM naturally handles "omissive causation". The CSM simulates what would have happened if the event of interest *had* taken place, and then comparing that counterfactual outcome to what actually happened. For example, when asked whether ball B went through the gate because ball A didn't hit it, the CSM simulates what would have happened if the collision had taken place, and how likely the outcome would then have been different.

In omissive causation, there is no cause event. In the case of physical support, there aren't any events at all – at least no events in the psychological sense where events mark changes of state (Glymour et al., 2010; Lewis, 1986a; Zacks & Tversky, 2001). Nothing changes in a stable block tower, it just sits there . Again, the CSM naturally extends to this type of causation that we may call "sustaining causation" (see Ross & Woodward, 2021, for relevant work in philosophy). A sustaining cause brings about an effect due to its continuing presence. In our case, an individual block in a tower is a sustaining cause of the tower's stability. Sustaining causation reveals itself by the counterfactual simulation of what would have happened if the sustaining cause had been removed. In other words, a block sustains the tower's stability because the tower would collapse if the block were removed.

The CSM provides a principled and general framework for understanding people's causal judgments across a variety of types of causal relationships that include "event causation" (Gerstenberg, Goodman, et al., 2021), "causation by omission" (Gerstenberg & Stephan, 2021), and "sustaining causation". Psychological theories that rely on events to explain causal judgments have trouble with causation by omission, and don't apply to instances of sustaining causation such as physical support (e.g. Wolff, 2007; Wolff, Barbey, & Hausknecht, 2010). The CSM assumes that people build a mental model of the world, and that different kinds of causal judgments can all be understood as counterfactual operations on this mental model. More work is required to better understand the cognitive processes that underlie causal judgments according to the CSM. In the remainder of the discussion, we will highlight some limitations of the CSM as it applies to capturing judgments of physical support, and suggest directions for future research.

The nature of mental simulation

The CSM assumes that people judge an object to be the cause of stability by mentally simulating what would happen if that object was not there. To do so, the CSM employs a physics engine for representing the scene and for simulating what would happen (Smith et al., under review; Ullman et al., 2017). To capture the gradedness in participants' judgments, the CSM incorporates different sources of uncertainty including *perceptual uncertainty* about the position of the blocks, *intervention uncertainty* about the removal of the block, and *dynamic uncertainty* about how the scene would unfold. With these sources of uncertainty, the CSM accurately captures participants' judgments.

The fact that these sources of uncertainty are sufficient for capturing participants' judgments does of course not mean that they are necessary. It's very plausible that there are other aspects of the scene that participants are uncertain about, and that alternative noise models would also accurately capture participants' judgments. For example, participants may be unsure about the degree of friction between the blocks, or the coefficient of restitution which determines how elastic the collisions are. It's also possible that participants consider a different counterfactual intervention from the one that the CSM implements. For example, instead of imagining what would happen if the block weren't there, they might imagine what would happen if the block was perturbed (without removing it). Although the CSM captures participants' judgments well, we do not claim that people are running counterfactual simulations in exactly that way. We do believe, however, that mental simulation, in some form, is critical for understanding physical support. The fact that the features model failed to capture participants' judgments as well as the CSM did, lends some support for this claim. Recent work has demonstrated that even primates engage in mental simulation in a physical prediction task (Rajalingham, Piccato, & Jazayeri, 2021).

The counterfactual simulations that are required to assess what would happen to the towers if the black block were removed are fairly complex. We are not arguing that mental simulation must act on a perfectly faithful representation of the world – there are clear limitations to what aspects of a scene people can represent and simulate. Ludwin-Peery et al. (2021) argue that people display errors in physical judgments that are at odds with simulation over a full representation of complex scenes. For instance, they show that when people choose one of four possible end states of a falling tower, they will often choose scenes in which one block has been removed or added. If people were representing each block individually, like a physics engine does, then such errors should have been unlikely. We believe that it's plausible that people construct a simplified mental representation of the scene, and that they then run mental simulations over that representation. Future work is required to better understand how people combine what they know about the physical world with what they see in a particular situation, to build a mental representation of the scene that is tailored to the task at hand (Ullman et al., 2017).

A related question is whether physical support needs to be inferred (via mental simulation), or whether it can be directly perceived (Little & Firestone, 2021). There is a rich literature on the perception of causality. Simple events, such as the classic Michottean launching event (Michotte, 1946/1963), look causal to us and it doesn't feel like we're engaging in mental simulation when judging that the first ball caused the second ball to move. What role physical knowledge plays in such simple situations is disputed (Bechlivanidis, Schlottmann, & Lagnado, 2019; Kominsky et al., 2017). Judgments about more complex cases, however, feel more inferential. For example, when judging whether one ball caused another ball to go into a hole, people spontaneously simulate what would have happened in the relevant counterfactual situation (Gerstenberg et al., 2017).

Where does physical support stand? Sometimes, it feels like we can directly see support relations. We see that the legs support the table top, and that the pillars support the roof. In more complex cases like the ones that we consider in this paper, it feels more like we have to engage our inferential abilities to make judgments about support (see Pramod, Cohen, Tenenbaum, & Kanwisher, 2021, for evidence of abstract representations of physical stability in the brain). It is plausible that in our experiments, participants relied on a combination of more general visual features and mental simulation. Smith et al. (2013) have shown that participants' predictions about whether a moving ball will first hit a red or green patch in a maze are generally well-accounted for by a noisy simulation model. However, there were also some trials in which participants' responded more rapidly than the model did. For example, participants realized that when the red patch was outside of an area that physically contained the ball and the green patch, the ball had to eventually hit the green patch (and it was impossible to hit the red patch). So participants seemed to draw both on more general topological information (such as containment) to make rapid inferences about what was possible, as well as on mental simulation to make predictions about what was probable (see also Smith, de Peres, Vul, & Tenenbaum, 2017). In a similar way, people may have learned visual shortcuts that are reliable predictors for stability in many situations.

Future work is required to delineate more clearly how visual features and mental simulation contribute to participants' physical predictions. We focused here on judgments of physical support but hypothetical simulations are also critical for planning and decision-making (Allen et al., 2020; Bapst et al., 2019; Baradel et al., 2019; Hamrick et al., 2018; Yildirim et al., 2017). For example, when deciding which block to pick in Jenga, a player needs to mentally simulate what will happen. Future work may ask people to take actions so as to make a tower more robust, or, conversely, to remove an object that is most likely to make the tower fall.

Supporting versus preventing from falling

In our experiments we didn't ask participants directly about physical support. Instead we asked them to judge how responsible one block was for the other blocks staying on the table (Experiments 1 and 2), or for one particular block (Experiment 3). We conceptualized the notion of responsibility as preventing from falling. A block is responsible to the extent that removing it from the scene would result in the other blocks falling from the table. While physical support and prevention from falling (or moving) often go together, there are also situations where these two notions intuitively come apart.

Figure 15a and b show two examples in which the black block is responsible for the white block staying on the table. In both situations, the white block would fall off the table if the black block weren't there. However, while it feels right in Figure 15a to say that the black block supports the white block, it doesn't feel quite right to say so in Figure 15b. “Physically supporting” is subtly different from “preventing from falling”. The Oxford Dictionary defines “to support” as “bear all or part of the weight of; hold up”, and give the

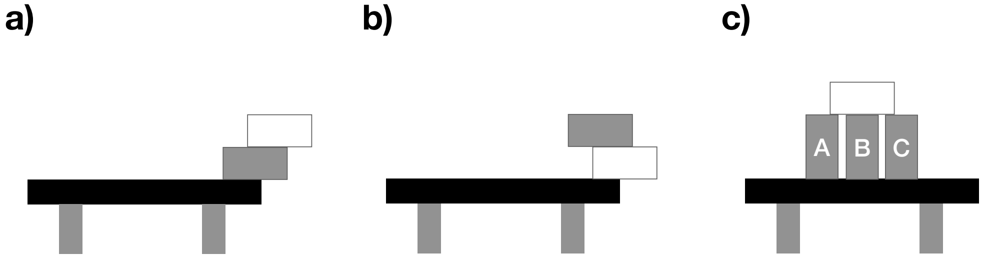


Figure 15. Example scenes. In a) and b), the black block is responsible for the white block staying on the table. The white block would fall off the table if the black block weren't there. However, only in a) but not in b) does it seem appropriate to say that the black block supports the white block, suggesting that the notion of "physically supporting" is different from the notion of "preventing from falling". In c) three black blocks are jointly responsible for the white block not moving. However, even if block A hadn't been there, the white block still wouldn't have moved. Only if at least two of the black blocks were removed, would the white block move.

example of "the dome was supported by a hundred white columns". So it seems like for one object to support another, it needs to be positioned underneath. Future work is required to say more about how exactly the notions of responsibility for stability, prevention from falling (or moving), and physical support are related to one another.

Overdetermination

The CSM captures people's responsibility judgments by considering what would happen if the target object were removed. The black object is responsible for the white object if the white object would fall without the black object. However, it's easy to conceive of situations in which removing the black object wouldn't make the white object fall but where we nevertheless feel that the black object carries some responsibility for the white object's stability. Consider the situation in Figure 15c. The three black blocks A, B, and C support the white block that rests on top of them. To what extent is each block responsible for the white block's stability (not considering here whether it would fall off the table, but instead whether it would fall at all)? Intuitively, each of the black blocks is somewhat responsible for the white block's stability. However, none of the blocks individually makes a difference. Even if block A hadn't been there, the white block would still have stayed exactly where it is.

In the literature on causation, a scenario like the one depicted in Figure 15c is known as an instance of overdetermination (Gerstenberg, Goodman, et al., 2021; Gerstenberg & Lagnado, 2010; Lagnado et al., 2013; Paul & Hall, 2013). Cases of overdetermination trouble theories that aim to explain causal relationships in terms of simple counterfactual dependence. To deal with such situations, counterfactual theories have been expanded to consider not only whether the candidate cause would have made a difference in the actual situation, but also whether it could have made a difference in another possible situation (Halpern, 2016; Halpern & Pearl, 2005; Woodward, 2003). For example, block A would make a difference to the white block's stability in a situation in which either block B or block C had been removed. Based on this idea, Chockler and Halpern (2004) developed a model according to which responsibility reduces the greater the distance is between the actual situation and a situation in which the candidate cause would have made a difference to the outcome (where distance is defined in terms of the number of variables whose values would need to be changed). So block A would still be responsible for supporting the white block to some degree because if either block B or block C wouldn't have been there, then block A would have been pivotal.

The current version of the CSM might assign some responsibility to each of the black blocks because it's possible that the white block would fall off due to different sources of simulation noise. Another way to capture the fact that each of the black blocks carries some responsibility is by imagining that external forces might perturb the scene. For example, in the current setup, the white block is likely to stay supported even if the table was bumped. But if one of the black blocks were removed then it would be more likely that a bump to the table would topple the white block over. Here again, the extent to which a block is responsible for another would not just be a function of the actual situation, but also take into account whether it would make a difference in other possible situations (see Grinfeld, Lagnado, Gerstenberg, Woodward, & Usher, 2020; Lewis, 1986b; Vasilyeva, Blanchard, & Lombrozo, 2018; Woodward, 2006). More research is needed to better understand how people assign responsibility in such cases of overdetermined stability.

Conclusion

Humans have a remarkable grasp on the physical world. We believe that this understanding is achieved by building mental models of the world that support the simulation of counterfactual possibilities. The counterfactual simulation model captures people's causal judgments about events (Gerstenberg, Goodman, et al., 2021) and omissions (Gerstenberg & Stephan, 2021), and it also naturally handles people's judgments about physical support. A block supports a tower by preventing it from falling. While most existing causal theories only apply to event causation, the CSM provides a unifying framework that explains people's causal judgments across a variety of different kinds of causal relationships.

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

Allen, K. R., Smith, K. A., & Tenenbaum, J. B. (2020). Rapid trial-and-error learning with simulation supports flexible tool use and physical reasoning. *Proceedings of the National Academy of Sciences*, *117*(47), 29302–29310.

Bapst, V., Sanchez-Gonzalez, A., Doersch, C., Stachenfeld, K., Kohli, P., Battaglia, P., & Hamrick, J. (2019). Structured agents for physical construction. In *International conference on machine learning* (pp. 464–474).

Baradel, F., Neverova, N., Mille, J., Mori, G., & Wolf, C. (2019). Cophy: Counterfactual learning of physical dynamics. *arXiv preprint arXiv:1909.12000*.

Battaglia, P., Pascanu, R., Lai, M., Rezende, D. J., et al. (2016). Interaction networks for learning about objects, relations and physics. In *Advances in neural information processing systems* (pp. 4502–4510).

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

Bear, D. M., Wang, E., Mrowca, D., Binder, F. J., Tung, H.-Y. F., Pramod, R., ... others (2021). Physion: Evaluating physical prediction from vision in humans and machines. *arXiv preprint arXiv:2106.08261*.

Bechlivanidis, C., Schlottmann, A., & Lagnado, D. A. (2019, April). Causation without realism. *Journal of Experimental Psychology: General*. doi: 10.1037/xge0000602

Beck, S. R. (2015). Why what is counterfactual really matters: A response to Weisberg and Gopnik (2013). *Cognitive Science*, *40*(1), 253–256. Retrieved from `https://doi.org/10.1111%2Fcogs.12235` doi: 10.1111/cogs.12235

Beebee, H. (2004). Causing and nothingness. In J. Collins, N. Hall, & L. A. Paul (Eds.), *Causation and counterfactuals* (pp. 291–308). MA: MIT Press Cambridge.

Beller, A., Bennett, E., & Gerstenberg, T. (2020). The language of causation. In *Proceedings of the 42nd Annual Conference of the Cognitive Science Society.*

Bramley, N. R., Gerstenberg, T., Tenenbaum, J. B., & Gureckis, T. M. (2018). Intuitive experimentation in the physical world. *Cognitive Psychology*, *105*, 9–38.

Chang, M. B., Ullman, T., Torralba, A., & Tenenbaum, J. B. (2017). A compositional object-based approach to learning physical dynamics. In *International conference on learning representations.*

Chockler, H., & Halpern, J. Y. (2004). Responsibility and blame: A structural-model approach. *Journal of Artificial Intelligence Research*, *22*(1), 93–115.

Cortesa, C. S., Jones, J. D., Hager, G. D., Khudanpur, S., Landau, B., & Shelton, A. L. (2018). Constraints and development in children's block construction. In *Cogsci.*

Craik, K. J. W. (1943). *The nature of explanation.* Cambridge, UK: Cambridge University Press.

Dowe, P. (2000). *Physical causation.* Cambridge, England: Cambridge University Press.

Fischer, J., Mikhael, J. G., Tenenbaum, J. B., & Kanwisher, N. (2016, aug). Functional neuroanatomy of intuitive physical inference. *Proceedings of the National Academy of Sciences*, *113*(34), E5072–E5081. Retrieved from `http://dx.doi.org/10.1073/pnas.1610344113` doi: 10.1073/pnas.1610344113

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. Retrieved from https://doi.org/10.1037%2F0096-3445.117.4.395 doi: 10.1037/0096-3445.117.4.395

Gerstenberg, T. (2022). What would have happened? counterfactuals, hypotheticals, and causal judgments. *PsyArXiv*. Retrieved from https://psyarxiv.com/rsb46

Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2021). A counterfactual simulation model of causal judgments for physical events. *Psychological Review*, *128*(6), 936–975.

Gerstenberg, T., & Icard, T. F. (2020). Expectations affect physical causation judgments. *Journal of Experimental Psychology: General*, *149*(3), 599–607.

Gerstenberg, T., & Lagnado, D. A. (2010). Spreading the blame: The allocation of responsibility amongst multiple agents. *Cognition*, *115*(1), 166–171.

Gerstenberg, T., Peterson, M. F., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2017, oct). Eye-tracking causality. *Psychological Science*, *28*(12), 1731–1744. Retrieved from https://doi.org/10.1177%2F0956797617713053 doi: 10.1177/0956797617713053

Gerstenberg, T., Siegel, M. H., & Tenenbaum, J. B. (2021). What happened? reconstructing the past from vision and sound. *PsyArXiv*. Retrieved from https://psyarxiv.com/tfjdk

Gerstenberg, T., & Stephan, S. (2021). A counterfactual simulation model of causation by omission. *Cognition*. Retrieved from https://psyarxiv.com/wmh4c/

Gerstenberg, T., & Tenenbaum, J. B. (2017). Intuitive theories. In M. Waldmannn (Ed.), *Oxford handbook of causal reasoning* (pp. 515–548). Oxford University Press.

Glymour, C., Danks, D., Glymour, B., Eberhardt, F., Ramsey, J., Scheines, R., … Zhang, J. (2010). Actual causation: a stone soup essay. *Synthese*, *175*(2), 169–192.

Grinfeld, G., Lagnado, D., Gerstenberg, T., Woodward, J. F., & Usher, M. (2020). Causal responsibility and robust causation. *Frontiers in Psychology*, *11*, 1069. Retrieved from https://www.frontiersin.org/article/10.3389/fpsyg.2020.01069 doi: 10.3389/fpsyg.2020.01069

Groth, O., Fuchs, F. B., Posner, I., & Vedaldi, A. (2018). Shapestacks: Learning vision-based physical intuition for generalised object stacking. In *Proceedings of the european conference on computer vision (eccv)* (pp. 702–717).

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.

Gweon, H., Asaba, M., & Bennett-Pierre, G. (2017). Reverse-engineering the process: Adults' and preschoolers' ability to infer the difficulty of novel tasks. In *Cogsci.*

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

Halpern, J. Y., & Pearl, J. (2005). Causes and explanations: A structural-model approach. Part I: Causes. *The British Journal for the Philosophy of Science*, *56*(4), 843–887.

Hamrick, J. B., Allen, K. R., Bapst, V., Zhu, T., McKee, K. R., Tenenbaum, J. B., & Battaglia, P. W. (2018). *Relational inductive bias for physical construction in humans and machines.*

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.

Henne, P., Pinillos, Á., & De Brigard, F. (2017). Cause by omission and norm: Not watering plants. *Australasian Journal of Philosophy*, *95*(2), 270–283.

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.

Hume, D. (1748/1975). *An enquiry concerning human understanding.* Oxford University Press.

Janner, M., Levine, S., Freeman, W. T., Tenenbaum, J. B., Finn, C., & Wu, J. (2019). Reasoning about physical interactions with object-centric models. In *International conference on learning representations* (pp. 1–12).

Kominsky, J. F., Strickland, B., Wertz, A. E., Elsner, C., Wynn, K., & Keil, F. C. (2017). Categories and constraints in causal perception. *Psychological Science*, *28*(11), 1649–1662.

Kubricht, J. R., Holyoak, K. J., & Lu, H. (2017, oct). Intuitive physics: Current research and controversies. *Trends in Cognitive Sciences*, *21*(10), 749–759. Retrieved from `https://doi.org/10.1016%2Fj.tics.2017.06.002` doi: 10.1016/j.tics.2017.06.002

Lagnado, D. A., Gerstenberg, T., & Zultan, R. (2013). Causal responsibility and counterfactuals. *Cognitive Science*, *47*, 1036–1073.

Langenhoff, A. F., Wiegmann, A., Halpern, J. Y., Tenenbaum, J. B., & Gerstenberg, T. (2021). Predicting responsibility judgments from dispositional inferences and causal attributions. *Cognitive Psychology*, *129*, 101412.

Lerer, A., Gross, S., & Fergus, R. (2016). Learning physical intuition of block towers by example. *arXiv preprint arXiv:1603.01312*.

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

Lewis, D. (1986a). Events. In *Philosophical papers* (Vol. II, pp. 241–270). Oxford University Press.

Lewis, D. (1986b). Postscript C to 'Causation': (Insensitive causation). In *Philosophical papers* (Vol. 2). Oxford: Oxford University Press.

Little, P. C., & Firestone, C. (2021). Physically implied surfaces. *Psychological Science*, *32*(5), 799–808.

Livengood, J., & Machery, E. (2007). The folk probably don't think what you think they think: Experiments on causation by absence. *Midwest Studies in Philosophy*, *31*(1), 107–127.

Ludwin-Peery, E., Bramley, N. R., Davis, E., & Gureckis, T. M. (2021). Limits on simulation approaches in intuitive physics. *Cognitive Psychology*, *127*, 101396.

Mackie, J. L. (1974). *The cement of the universe.* Oxford: Clarendon Press.

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

McCloskey, M., Caramazza, A., & Green, B. (1980). Curvilinear motion in the absence of external forces: Na´ıve beliefs about the motion of objects. *Science*, *210*(4474), 1138–1141.

McCloskey, M., Washburn, A., & Felch, L. (1983). Intuitive physics: The straight-down belief and its origin. *Journal of Experimental Psychology: Learning, Memory, and*

Cognition, *9*(4), 636–649.

McGrath, S. (2005). Causation by omission: A dilemma. *Philosophical Studies*, *123*(1), 125–148.

Michotte, A. (1946/1963). *The perception of causality.* Basic Books.

Mitko, A., & Fischer, J. (2020). When it all falls down: the relationship between intuitive physics and spatial cognition. *Cognitive research: principles and implications*, *5*, 1–13.

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.

Pramod, R., Cohen, M. A., Tenenbaum, J. B., & Kanwisher, N. G. (2021). Invariant representation of physical stability in the human brain. *bioRxiv*.

Rajalingham, R., Piccato, A., & Jazayeri, M. (2021). The role of mental simulation in primate physical inference abilities. *bioRxiv*. Retrieved from `https://www.biorxiv.org/content/early/2021/01/17/2021.01.14.426741` doi: 10.1101/2021.01.14.426741

Ross, L. N., & Woodward, J. (2021). Irreversible (one-hit) and reversible (sustaining) causation.

Salmon, W. C. (1994). Causality without counterfactuals. *Philosophy of Science*, *61*(2), 297–312.

Sanborn, A. N., & Chater, N. (2016). Bayesian brains without probabilities. *Trends in Cognitive Sciences*, *20*(12), 883–893.

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.

Schaffer, J. (2016). The metaphysics of causation. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Fall 2016 ed.). Metaphysics Research Lab, Stanford University. `https://plato.stanford.edu/archives/fall2016/entries/causation-metaphysics/`.

Schwartz, D. L., & Black, T. (1999). Inferences through imagined actions: Knowing by simulated doing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(1), 116–136.

Sloman, S. A. (2005). *Causal models: How people think about the world and its alternatives.* Oxford University Press, USA.

Smith, K. A., Dechter, E., Tenenbaum, J., & Vul, E. (2013). Physical predictions over time. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th annual meeting of the cognitive science society* (pp. 1342–1347).

Smith, K. A., de Peres, F. A. B., Vul, E., & Tenenbaum, J. B. (2017). Thinking inside the box: Motion prediction in contained spaces using simulation. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. Davelaar (Eds.), *Proceedings of the 39th Annual Conference of the Cognitive Science Society* (pp. 3209–3214). Austin, TX: Cognitive Science Society.

Smith, K. A., Hamrick, J. B., Sanborn, A. N., Battaglia, P. W., Gerstenberg, T., Ullman, T. D., & Tenenbaum, J. B. (under review). Probabilistic models of physical reasoning. In T. L. Griffiths, N. Chater, & J. B. Tenenbaum (Eds.), *Reverse engineering the mind:*

Probabilistic models of cognition.

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

Smith, K. A., & Vul, E. (2014). Looking forwards and backwards: Similarities and differences in prediction and retrodiction. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th Annual Conference of the Cognitive Science Society* (pp. 1467–1472). Austin, TX: Cognitive Science Society.

Suppes, P. (1970). *A probabilistic theory of causation.* Amsterdam: North-Holland.

Ullman, T. D., Spelke, E., Battaglia, P., & Tenenbaum, J. B. (2017, sep). Mind games: Game engines as an architecture for intuitive physics. *Trends in Cognitive Sciences*, *21*(9), 649–665. Retrieved from `https://doi.org/10.1016%2Fj.tics.2017.05.012` doi: 10.1016/j.tics.2017.05.012

Ullman, T. D., Stuhlmüller, A., Goodman, N. D., & Tenenbaum, J. B. (2018). Learning physical parameters from dynamic scenes. *Cognitive Psychology*, *104*, 57–82.

Ullman, T. D., & Tenenbaum, J. B. (2020). Bayesian models of conceptual development: Learning as building models of the world. *Annual Review of Developmental Psychology*, *2*, 533–558.

Vasilyeva, N., Blanchard, T., & Lombrozo, T. (2018, Apr). Stable causal relationships are better causal relationships. *Cognitive Science*, *42*(4), 1265–1296. Retrieved from `http://dx.doi.org/10.1111/cogs.12605` doi: 10.1111/cogs.12605

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

Wolff, P., Barbey, A. K., & Hausknecht, M. (2010). For want of a nail: How absences cause events. *Journal of Experimental Psychology: General*, *139*(2), 191–221.

Woodward, J. (2003). *Making things happen: A theory of causal explanation.* Oxford, England: Oxford University Press.

Woodward, J. (2006). Sensitive and insensitive causation. *The Philosophical Review*, *115*(1), 1–50.

Wu, J., Yildirim, I., Lim, J. J., Freeman, B., & Tenenbaum, J. (2015). Galileo: Perceiving physical object properties by integrating a physics engine with deep learning. In *Advances in neural information processing systems* (pp. 127–135).

Yildirim, I., Gerstenberg, T., Saeed, B., Toussant, M., & Tenenbaum, J. B. (2017). Physical problem solving: Joint planning with symbolic, geometric, and dynamic constraints. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. Davelaar (Eds.), *Proceedings of the 39th Annual Conference of the Cognitive Science Society* (pp. 3584–3589). Austin, TX: Cognitive Science Society.

Yildirim, I., Saeed, B., Bennett-Pierre, G., Gerstenberg, T., Tenenbaum, J. B., & Gweon, H. (2019). Explaining intuitive difficulty judgments by modeling physical effort and risk. *Proceedings of the 41st Annual Conference of the Cognitive Science Society.*

Zacks, J. M., & Tversky, B. (2001). Event structure in perception and conception. *Psychological bulletin*, *127*(1), 3–21.

Zultan, R., Gerstenberg, T., & Lagnado, D. A. (2012). Finding fault: Counterfactuals and causality in group attributions. *Cognition*, *125*(3), 429–440.

Appendix

Table A1

Correlation coefficients between individual features and participants' responsibility judgments for each of the three experiments. Note: *The* scene features*,* black block features*,* other block features*, and* all features *columns show how well regressions that combine these features correlate with participants' responsibility judgments.* other block *refers to the white block in Experiment 3. See Table 1 for a description of each feature. The table shows that which features work best differs between the experiments. Scene features are most important for Experiment 1, black block features for Experiment 2, and other block features (i.e. the features of the white block) for Experiment 3.*

	Experiment 1	Experiment 2	Experiment 3	All
n_blocks	.13	−.03	.12	.05
avg_edge_dist	.37	−.08	−.04	.03
avg_angle	−.15	.00	-.04	−.05
avg_y	.71	.07	.09	.28
scene features	**.77**	**.15**	**.18**	**.31**
black_y	−.21	−.74	−.27	−.64
black_edge_dist	−.04	.12	−.26	.07
black_angle	.04	−.05	−.30	−.03
black_above	.55	.84	.14	.70
black block features	**.58**	**.87**	**.43**	**.75**
other_y	-	-	.64	-
other_edge_dist	-	-	−.01	-
other_angle	-	-	.20	-
other_black_pile	-	-	-	-
other block features	**-**	**-**	**.75**	**-**
all features	**.84**	**.89**	**.83**	**.80**