Mental Jenga

A counterfactual simulation model of causal judgments about 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

Author Note

*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`

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 causal judgments about 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 by running noisy simulations that incorporate different sources of uncertainty. Participants' responsibility judgments are closely related to counterfactual predictions: a block is more responsible when many other blocks would fall if it were removed. By construing physical support as preventing from falling, the CSM provides a unified account of how causal judgments in dynamic and static physical scenes arise from the process of counterfactual simulation.

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

Mental Jenga

A counterfactual simulation model of causal judgments about physical support

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?

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 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 how responsible one object is for the stability of others. 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 a 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 various 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 particular block 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 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 closely relate 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 features of the actual situation, 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 limitations of the CSM and future challenges that lie ahead.

People's intuitive understanding of the physical world

People generally have a good sense of how the physical world works. We catch balls, stack stones, ride bikes, and build towers (Figure 1, top). While earlier work documented how people's physical intuitions sometimes fail (McCloskey, 1983; McCloskey, Caramazza, & Green, 1980; McCloskey, Washburn, & Felch, 1983), more recent work emphasizes situations in which they succeed (Kubricht, Holyoak, & Lu, 2017; Smith et al., in press). People can predict whether a tower will topple over (Battaglia, Hamrick, & Tenenbaum, 2013), or where a moving object will go next (Smith, Dechter, Tenenbaum, & Vul, 2013; Smith & Vul, 2013). They can also use their intuitive physical understanding to infer what happened. They can figure out where a ball was behind an occluder before it appeared (Smith & Vul, 2014), in which hole of a box a ball was dropped based on visual and auditory cues (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 infer unobservable physical properties such as the mass of different objects based on how they collide with one another (Sanborn,

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; a child stacking toy blocks. **Bottom**: The process of removing a block from a tower in the *Jenga* game. In the game *Jenga*, the goal is to remove a block from a stable tower and to put it on top without making the tower fall (unlike what happened here).

Mansinghka, & Griffiths, 2013) or based on how they form stable configurations of block towers (Hamrick, Battaglia, Griffiths, & Tenenbaum, 2016). People give causal explanations of what happened by comparing what actually happened with what would have happened if the candidate cause hadn't been present in the scene (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., in press; 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. They assume 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 perceptual uncertainty about where exactly the different blocks are located and 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 of future outcomes. This distribution is then used to capture people's graded judgments. For example, consider a stable, well-supported tower A compared to 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. This approach of modeling people's intuitive understanding of the physical world is sometimes referred to as "noisy Newtons": "noisy" because noise is added to the simulations to capture people's uncertainty, and "Newtons" because the dynamics of the physics engine approximates

Newtonian dynamics (Smith et al., in press).

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 various ways for how people might learn to make predictions about block towers and 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., in press; 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?). Predictions and explanations operate on the same mental model but require distinct computations (Gerstenberg & Tenenbaum, 2017). For prediction, one only needs to unroll a simulation of what will happen forward. Giving causal explanations, however, involves a comparison of what actually happened with what would have happened otherwise (Gerstenberg, in press).

But are such counterfactual comparisons really necessary? Maybe it's sufficient to just focus on what's true of the actual situation? In the philosophical literature, there are two major theoretical frameworks for thinking about causation. According to process theories, causation is understood as 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 if 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 if B wouldn't have moved had A been removed from the scene. While process theories ground causation solely in terms of what actually happened, dependence theories rely on a comparison between what actually happened and what would have happened otherwise.

Drawing on both theoretical frameworks, Gerstenberg, Goodman, et al. (2021) developed the *counterfactual simulation model* (CSM) of causal judgments about dynamic physical events. In line with dependence theories, the CSM assumes that judging whether one billiard ball A caused another 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 have been different without ball A, the more they judge that ball A caused the outcome. Gerstenberg et al. (2017) showed that people spontaneously engage in counterfactual simulation when making causal judgments as evidenced by their eye movements. People don't just look at what actually happened, they look at where ball B would have gone if ball A hadn't been present in the scene (see also Gerstenberg, in press).

Sustaining causation and physical support. Gerstenberg, Goodman, et al. (2021) asked participants to judge whether something caused an outcome or prevented it from happening. Gerstenberg and Stephan (2021) used the CSM to explain participants' causal judgments about omissions. They showed participants video clips in which ball B went through a gate (or missed the gate), while ball A was just lying still in the corner of the scene. Participants judged whether ball B went through the gate (or missed the gate) because ball A didn't hit it. As predicted by the CSM, participants' causal judgments increased the more likely it was that the outcome in the counterfactual situation (in which ball A had hit ball B) would have been different from what actually happened.

Here, we build on the CSM and apply it to understanding people's judgments of physical support. Judging physical support is closely related to judging causation. Table 1 categorizes different types of causal relationships based on the presence or absence of the cause and effect events. Gerstenberg, Goodman, et al. (2021) focused on situations in which the cause event was present and it either prevented the effect event (absent) or caused it to happen (present). Gerstenberg and Stephan (2021) looked at situations in which the cause event was absent and the effect event was present. In this paper, we fill in the table's missing cell: situations in which there is neither a cause event nor an effect event. We will call this type of causal relationship "sustaining causation".

In our experiments, participants view images of static scenes depicting a tower of

Table 1
Different types of causal relationships based on the absence or presence of cause events and effect events. The citations point to prior work that has used the CSM to explain these kinds of causal relationships.

		effect event	
		absent	present
cause event	absent	**sustaining causation**	**causation by omission** (Gerstenberg & Stephan, 2021)
	present	**prevention** (Gerstenberg, Goodman, et al., 2021)	**causation** (Gerstenberg, Goodman, et al., 2021)

blocks. In these images, there are no events – at least not in the psychological sense of events capturing state changes (Glymour et al., 2010; Lewis, 1986a; Schaffer, 2016; Zacks & Tversky, 2001). The tower just stands still and nothing is moving. Nevertheless, we may wonder how responsible a particular block is for another block's staying on the table on which the block tower rests. Is block A a sustaining cause of block B's staying on the table? The CSM answers this question by simulating what would happen if block A were removed from the scene. The more certain it is the that block B would fall off the table in that case, the more responsible block A is for block B's staying on the table. Another way of putting it is that a sustaining cause prevents an alternative outcome from happening. Block A is responsible to the extent that it prevents block B from falling off the table.

Process theories of causation most naturally apply to situations in which both the cause and effect events are present (i.e. the bottom right cell in Table 1). These theories generally struggle when absences are involved because an absence transfers no force (although see Wolff, Barbey, & Hausknecht, 2010). The case of sustaining causation is particularly troublesome for process theories. 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. Counterfactual theories, such as the CSM, apply more flexibly to the different types of causal relationships. According to the CSM, judging sustaining causation requires going beyond what can be directly perceived. To see whether counterfactual simulations are necessary for capturing people's judgments, we compare the CSM with an alternative model that relies exclusively on visual information that's present in the scene.

The notion of sustaining causation has received little interest in work on causal cognition so far. In philosophy, Ross and Woodward (2022) have discussed sustaining causation in the context of reversible causal relationships. Causal relationships are reversible when an earlier change can be undone at a later point in time. For example, a

light switch is a sustaining cause of the light. The light is on when the switch is on, and off when the switch is off. The light switch is also a reversible cause: one can go back and forth between the light's two different states by flipping the switch. In the block towers we focus on, there is sustaining causation without reversibility – one often cannot reconstruct a tower by putting the block back to where it was before because the tower may have collapsed.

While there are many different ways in which a cause can sustain a particular outcome, we focus here on physical support. That said, in our experiments, we don't ask participants directly about physical support. Instead, we ask them to judge the extent to which one block was responsible for other blocks (or one specific block) staying on a table on which the blocks are stacked. The CSM captures these responsibility judgments by simulating whether the presence of the block prevents the other blocks from falling off. "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 gives the example of "the dome was supported by a hundred white columns". In many situations, judgments about physical support and responsibility are likely to go together. However, they may also come apart. For example, it's possible that object A prevents object B from falling off the table even though object B is below (or to the side of) object A. In these situations, it wouldn't seem right to say that object A supports object B. Physical support seems to have two requirements: 1) A supports B when A plays a role in preventing B from falling (or moving), and 2) A is positioned underneath B. The model we develop below is a model of the first component of physical support. We return to the question of how physical support, responsibility, 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 approximate simulations of what would happen (Sanborn & Chater, 2016). We apply the CSM to block towers like the one shown in Figure 2. We probe the CSM in several ways, asking it to predict which blocks would fall if the black one were removed (*selection*), how many of the blocks would fall (*prediction*), and how responsible the black block is for the red blocks staying on the table (*responsibility*).

Scope of the model

Before describing in more detail how the CSM works, we first clarify its scope.

Block towers as a case study. Physical support is a broad concept: it applies to blocks supporting one another in a tower, but also to monitors on arms on tables, to the foundation and structure of a house, and to the stability of a surfer on a wave. Ultimately, we aim to better understand how people think about all the different ways in which physical support grounds out in the world. Here, our model focuses on a constrained setting: physical support in the context of block towers (see Figure 2).

We believe that the core foundations on which the CSM rests are relevant beyond the case study of block towers. The CSM makes three core assumptions about how people reach a causal judgment in a particular situation: 1) people represent the situation with a mental model, 2) they imagine a counterfactual intervention on that mental model, and 3) they mentally simulate what the consequences of this counterfactual intervention would be. On this computational-level description, we are committed to the CSM's core assumptions. In order to apply the CSM to any given situation, the three components have to be implemented – and there are many possible ways of doing so. We briefly motivate our choices below.

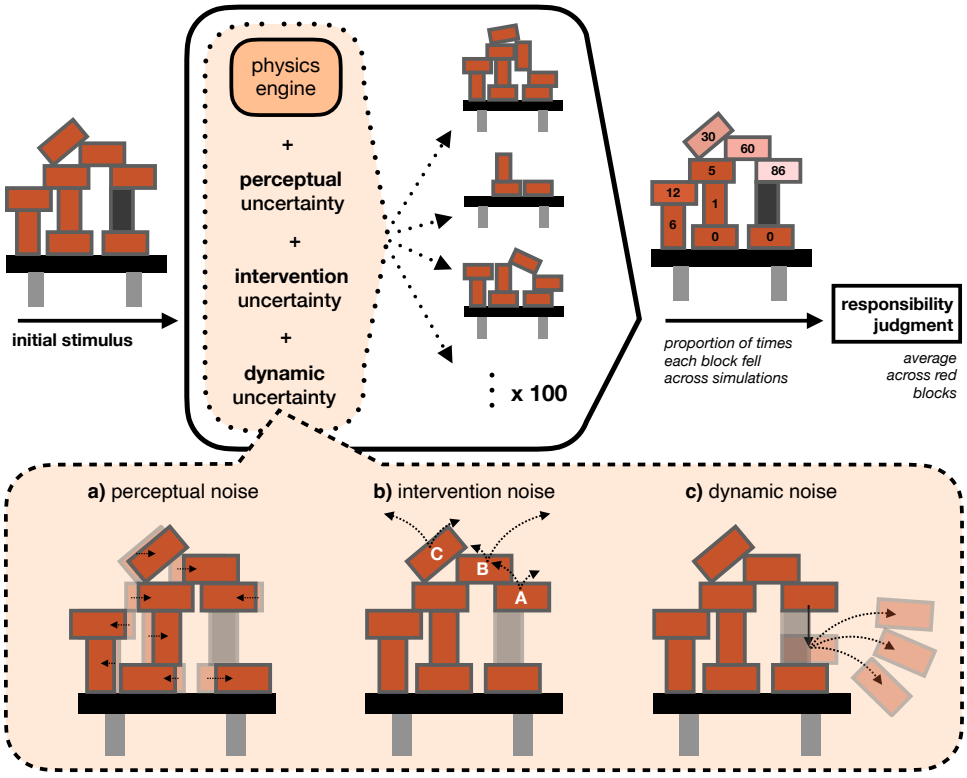

Figure 2. Schematic illustration of the *counterfactual simulation model* (CSM) applied to a block tower. Given a scene as input (top left), the CSM answers the question of how responsible the black block is for the red blocks staying on the table. It does so by simulating counterfactual rollouts of 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 model records which red blocks fell off the table. The CSM predicts 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 fall 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 determines how much noise is applied. You can play around with the model parameters and see the CSM in action here: `https://cicl-stanford.github.io/mental_jenga/interface`

Mental model. We assume that people's mental model of the situation is similar to the kinds of physics engines that are used in video games for simulating realistic physical

interactions (Gerstenberg & Tenenbaum, 2017; Smith et al., in press; Ullman et al., 2017). These game engines represent the physical scene as comprised of objects with attributes (such as friction and mass) and the physical forces that apply to these objects (such as gravity and collisions). 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 to the question of what would happen. However, people don't have access to this ground truth. Various sources of uncertainty 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, and the task at hand. While we assume that people construct a mental model of the scene we don't provide an answer to the question of how people learn these mental models (see Bramley et al., 2018; Rule, Tenenbaum, & Piantadosi, 2020; Ullman & Tenenbaum, 2020; Yi et al., 2019).

Using a physics engine as an approximation to people's mental model of the physical world is very helpful. It allows us to implement a concrete model that yields quantitative predictions that can be tested against people's judgments. That said, people's mental models may be quite different from physics engines. For example, it's possible that people don't explicitly represent all the objects in the scene (Ludwin-Peery et al., 2021), or that they represent the various kinds of forces that act upon the objects differently. We use the physics engine as a useful tool for implementing the idea that people build a mental model of the scene.

Counterfactual intervention. We assume that when people evaluate how responsible the black block is for the other blocks staying on the table, they consider to what extent the black block prevents the other blocks from falling off the table. Similarly, we assume that to assess prevention, people consider what would happen in the counterfactual situation in which the black block wasn't there. There are many possible

ways of construing such a counterfactual situation. For example, people could imagine that the black block just magically disappeared. Or they could imagine that it was pulled out of the tower in a certain way, similar to how one would remove a piece from a Jenga tower (see Figure 1). The CSM implements the counterfactual intervention in a way that accounts well for participants' judgments in our experiments. However, it's possible that people differ in what counterfactual comes to mind (Kominsky & Phillips, 2019; Phillips & Knobe, 2018).

When judging causation in dynamic scenes, people need to consider what *would have happened* if something in the past had been different (Gerstenberg, in press; Gerstenberg, Goodman, et al., 2021; Gerstenberg et al., 2017). 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 (see Gerstenberg, in press). To highlight the continuity with our prior work (Gerstenberg, Goodman, et al., 2021), we chose to call our model the counterfactual simulation model, rather than a hypothetical simulation model.

Mental simulation. We assume that people assess what would have happened in the relevant counterfactual situation by running a simulation in their mind (Gerstenberg & Tenenbaum, 2017; Kahneman & Tversky, 1982). Some of the most direct evidence for mental simulation comes from eye-tracking studies (Ahuja & Sheinberg, 2019; Beller, Xu, Linderman, & Gerstenberg, 2022; Crespi, Robino, Silva, & de'Sperati, 2012; Gerstenberg et al., 2017). For example, Gerstenberg et al. (2017) asked participants to judge whether ball

A caused ball B to go through a gate in a dynamic video clip. Participants' eye movements revealed that they didn't just look at the balls, they also looked at where ball B would have gone if ball A hadn't been present in the scene. Other evidence for mental simulation comes from studies in which simulation models accounted better for participants' judgments than alternative models that didn't rely on simulation (e.g. Battaglia et al., 2013; Gerstenberg, Siegel, & Tenenbaum, 2021; Rajalingham, Piccato, & Jazayeri, 2021; Smith et al., in press; Smith & Vul, 2013).

While simulation models are powerful tools for modeling people's judgments, they have limitations. For one, we don't know exactly what people's mental simulations actually look like. Ludwin-Peery et al. (2021) have shown that people make judgments about physical scenarios that violate the predictions of certain simulation models (see also Ludwin-Peery, Bramley, Davis, & Gureckis, 2020). For simulation models to yield graded predictions about people's judgments, the developers of these models make assumptions about possible sources of uncertainty in people's mental simulations. Often it's the case that these sources of uncertainty are underdetermined by the data. For example, uncertainty about what would happen if a block were removed from the tower could be modeled by adding uncertainty in how the object collisions play out, or in the amount of friction that's present between the objects. The CSM implements these mental simulations in a way that accounts well for people's judgments, but it's possible that the actual mental simulations that people run are quite different from that particular implementation.

We will now describe our implementation of the CSM. We first lay out how the CSM implements uncertainty about the counterfactual simulations. Then, we explain how the CSM uses counterfactual simulations to make predictions about the different kinds of judgments in our experiments.

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 are told 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 uncertainty is implemented in our model: the CSM takes the ground truth configuration of blocks (shown faintly in the background) and applies a small horizontal perturbation to each block's position, randomly moving some of 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. As a consequence, some of the scenes would no longer be stable after perceptual noise has been applied. To account for the fact that participants know that the initial scene is stable, we put all of the blocks to "sleep" after the perceptual noise was applied (see Ullman et al., 2017). A block that is asleep stays exactly where it is and only wakes up once another block collides with it.

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. We don't specify explicitly how this counterfactual state would come about. It's possible, for example, that some participants imagine that the black block simply disappeared while others imagine physically removing the block by pulling it out. We implemented intervention uncertainty by applying small, roughly upwards-directed impulses to red blocks located above the black block, mimicking the disturbance that would be caused if the black block were manually removed from the scene (Figure 2b).

First, the black block is removed from the scene by making it disappear. Then, a

random impulse is applied to all the blocks that were located above the black block. The angle of the impulse is the same for all the blocks, but the magnitude differs (see the dotted arrows in Figure 2b, showing the impulses applied to the red blocks in two possible simulations).. The angle is drawn uniformly from $[-\frac{\pi}{4}, \frac{\pi}{4}]$ around the vertically upwards direction. The magnitude of the impulse applied to each block is drawn independently from $\Gamma(\beta_i, 1)$ where $\Gamma(k, \theta)$ is a Gamma distribution with shape parameter k and scale parameter θ.

Overall, our implementation of intervention noise 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 an impulse is applied to each of them after the black block is removed.

Dynamic uncertainty. After the black block is removed, the physics engine simulates the dynamics of how the scene would unfold. Again, we assume that people have some uncertainty about how these dynamics would play out (Allen et al., 2020; Smith & Vul, 2013). The CSM models this dynamic uncertainty by adding noise to collisions as illustrated in Figure 2c. Each dotted arrow shows different samples of how the collision between the two blocks could produce different resulting trajectories. To model dynamic uncertainty, the CSM applies noise to the ground truth normal force that results from two objects coming into contact with one another. The model perturbs the direction of that force (without changing the magnitude): for a normal force expressed in polar coordinates as $\mathbf{F} = (F, \theta)$, we alter it so that $\mathbf{F'} = (F, \theta + \alpha)$ where $\alpha \sim \mathcal{N}(0, \beta_d)$.

We will show below that a 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 consider 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 noisy simulation model captures participants' judgments even better than the one we developed.

Because the model only includes a small subset of the potential sources of uncertainty that affect people's physical predictions, this subset will have to make up for any remaining uncertainty not modeled in our implementation. What this means in practice is that the degree of noise the model applies is likely exaggerated. For example, it's possible that the perceptual uncertainty that participants have about the exact location of each block is less than what the best-fitting model assumes. A model that included even more sources of uncertainty (such as uncertainty about object friction) may predict participants' judgments better with a lower degree of perceptual uncertainty.

Our goal is not so much to determine exactly which 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: one that we believe strikes a good balance between simplicity and complexity, and one that includes intuitively plausible sources of uncertainty. We employ model comparison techniques to justify this trade-off between simplicity and complexity. We 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 Experiments 1 and 2, 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 weren't there. In the *prediction condition*, participants indicated how many blocks would fall off the table. In the *responsibility*

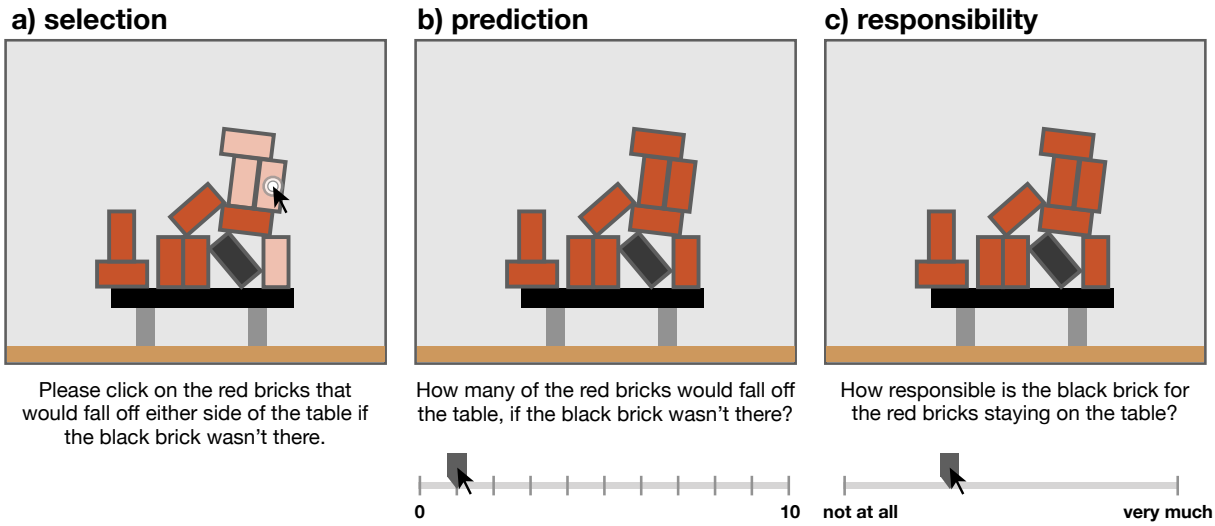

Figure 3. Schematic of the different experimental conditions in Experiments 1 and 2. Participants were asked to either **a)** select which blocks would fall if the black block wasn't there by clicking on the blocks (selection), **b)** judge on a slider how many blocks would fall (prediction), or **c)**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

condition, participants judged to what extent the black block was responsible for the red blocks staying on the table. In Experiment 3, participants were asked about the relationship between two specific blocks. In this experiment, participants either predicted whether the white block would fall off the table if the black block were removed (*prediction condition*), or judged to what extent the black block was responsible for the white block staying on the table (*responsibility condition*).

Figure 2 illustrates how the CSM yields graded predictions about how likely the different red blocks would fall off the table if the black block were removed. The model begins with an accurate encoding of the scene.[1] It then simulates the removal of the black block under different sources of uncertainty as described above. Each noisy simulation yields a potentially different result. For example, in one simulation a particular block may

[1]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 the addition of different types of noise, were performed with Box2D's engine. Further details about the implementation including are available online at `https://github.com/cicl-stanford/mental_jenga`

fall off the table, whereas in another simulation the 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).

To predict which blocks participants will select in the *selection condition*, the CSM uses the proportion of times with which each block fell off the table across the noisy simulations. 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' judgments in the *responsibility condition*, the CSM computes a linear mapping from the proportion of blocks that would fall off the table. For example, the black block would be more responsible if three out of four blocks were to fall off the table compared to five out of ten blocks.[2]

Features model

As an alternative to the CSM, we consider a *features model* that captures participants' judgments based on features of the scene. This model assumes that people's counterfactual judgments are not based on simulating what would happen to the blocks, but instead that they form judgments of whether blocks would fall using heuristics that rely on the current scene state. We implement this model by fitting a logistic regression from a collection of features to participants' predictions of how likely individual red blocks are to fall. This model 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 2 shows which 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 are present. *Black*

[2]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.

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.

Some of the features encode information about the vertical position of the blocks, their distance to the edge, and their rotation. Other features encode more higher-level information such as the total number of blocks in the scene, the number of red blocks that are above the black block (using the same definition of 'above' introduced earlier), and whether or not a red block was in the same pile of blocks as the black block (which we encoded by recursively checking whether a block makes contact with the black block, or with a block that makes contact with the black block, and so on). The features model thus

Table 2
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	`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
	`n_blocks`	total number of blocks, excluding the black block, in the tower
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

contains both low-level features as predictors as well as more abstract rules. Importantly, all of its features can be computed directly from the image. They don't encode any unobservable physical information (such as the forces the blocks exert on each other), or information that relies on running physical simulations of the scene.

When constructing the features model, we tried our best to find features that predicted participants' judgments. From a large set of initial features, we selected a subset of features using the following two criteria: a) the feature had to be sufficiently predictive in at least one of the experiments ($r > |0.1|$, see Table 3), and b) the feature didn't strongly correlate with any other feature ($r < |0.8|$, see Table A1).

Parameter fitting and model evaluation

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 would fall or stay on the table. Our implementation of 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 of the features model, we performed a logistic regression on the data. The features model has 13 free parameters: one for each of the features, plus one for the intercept. When discussing the results from individual experiments, we report the results of the CSM and the features model that were fitted to all three experiments. We will report Pearson's correlation (r) and root mean squared error (RMSE) as measures of

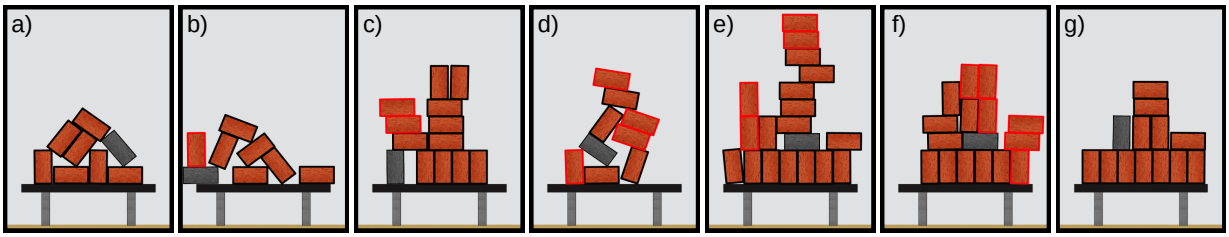

Figure 4. **Experiment 1**. Example stimuli. Red outlines indicate blocks that would fall off the table if the black block weren't there and black outlines indicate the blocks that would stay on the table. The outlines were not shown in the experiment.

fit to determine which model better accounted for the data.

In a separate section on 'Model comparison', we then report the results of cross-validation analyses that compare the features model, the full CSM, as well as lesioned versions of the CSM that only include a subset of the different sources of uncertainty. Cross-validation naturally handles the inherent trade-off between model complexity and fit to the data. We find that the full CSM best accounts for participants' responses.

Experiment 1: A wide array of block towers

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 of participants and the responsibility judgments from another group of participants. The greater the proportion of blocks 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 from MIT's institutional review board (COUHES #0812003014: Learning and Reasoning with Words and Concepts).

Participants. 121 participants ($M_{age} = 34$, $SD_{age} = 12$, 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 described below. No participant 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."[3] 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 answers 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).

We generated 42 towers of blocks that served as the stimuli for Experiment 1. To generate the stimuli, we randomly dropped 19 red blocks and 1 black block from above the

[3]In the experiments, we referred to the objects as "bricks" (rather than blocks). However, in this paper, we use the more generic term "blocks" unless we directly quote the questions that participants were asked in the experiments.

tabletop. As a result, some of the blocks would fall off the table, whereas others would remain on the table and settle into a stable configuration. We repeated this process many times and then selected 42 tower stimuli for the experiment using the following criteria: a) the black block was still on the table, b) the number of red blocks varied between the scenes (ranging from 4 to 19), and c) the number of red blocks that would fall off the table if the black block were removed varied between the scenes (ranging from 0 to 8).

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 = 2.1, range = 0 to 8). 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 as well as 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.

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 of each condition in turn, using these trials as illustrative examples.

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 show 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 features model does a better job matching the probabilities for the red blocks near the left edge. 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. Neither model captures participants' belief that the three blocks above the black block would almost certainly fall.

Figures 6a and 6d show how well the CSM and the features model capture participants' selections across all the red blocks in all of the trials. Table 3 shows how well

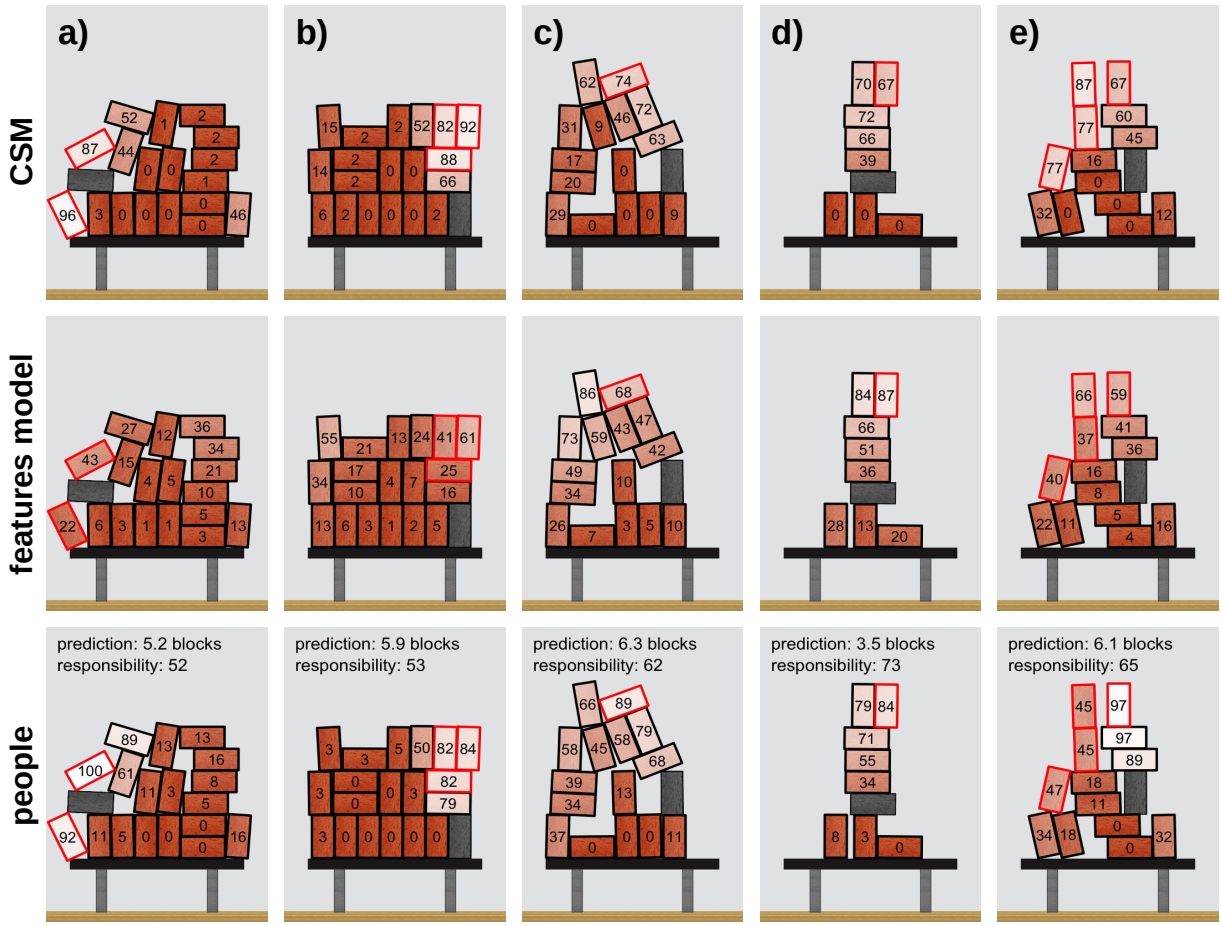

Figure 5. **Experiment 1**: Participants' selections of which blocks would fall (bottom row) together with the predictions of the *features model* (middle row) and the *counterfactual simulation model* (CSM, top row). The numbers on each block indicate the percentage of participants who thought that 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 red blocks to stay on the table. *Note*: The color fill gradient of the blocks maps onto 0 (red) and 100 (white). A bright red outline on a block indicates that a block would fall off the table according to ground truth. The outlines were not displayed in the experiment.

individual features and sets of features correlate with participants' selections. The y-position of a red block is a good predictor for whether a block will be selected by participants: blocks with a higher y-position are more likely to be selected.

Overall, the CSM does a better job of fitting participants' selections than the features

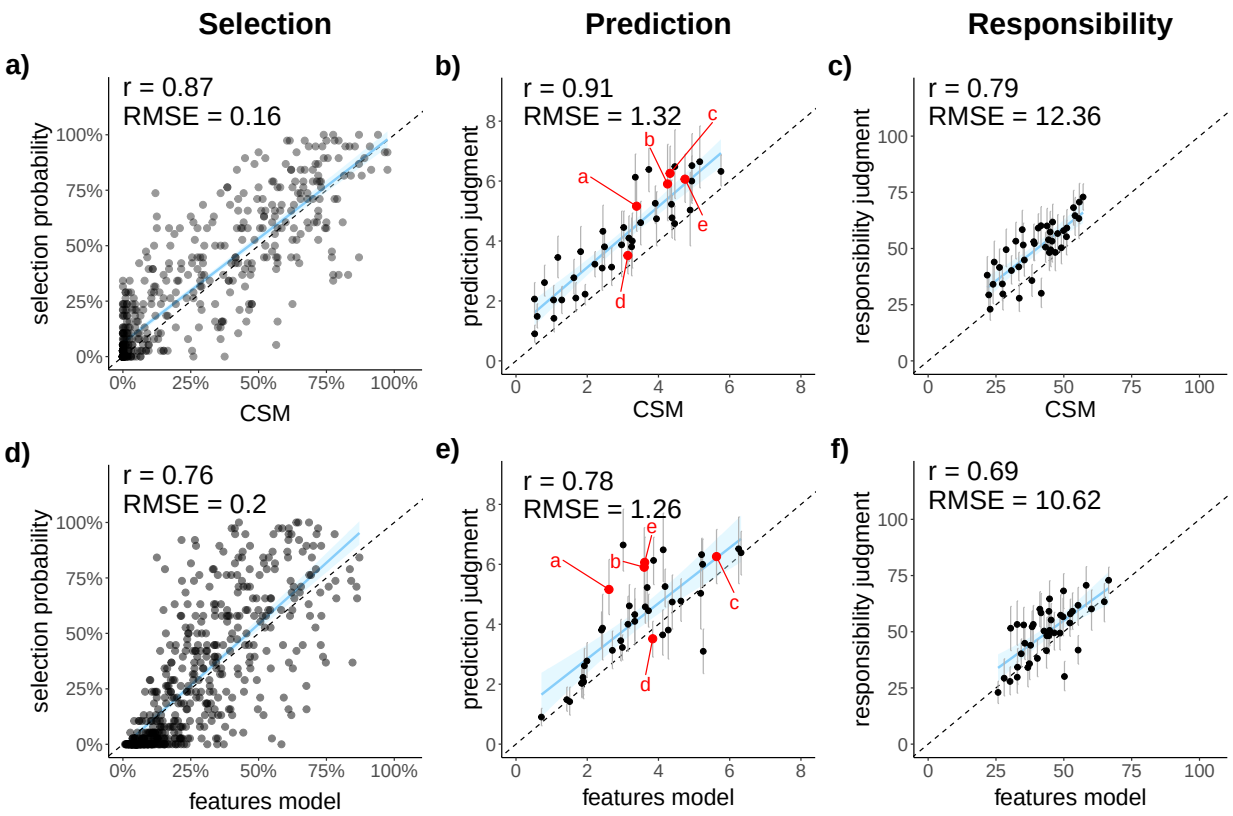

Figure 6. **Experiment 1**: Scatterplots showing the relationship between the CSM and participants' judgments at the top, and the relationship between the features model and participants' judgments at the bottom. The first column shows the results of the selection condition. Here, each data point represents 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 of the prediction condition. Here, each data point represents the average number of blocks that were predicted to fall in each trial. The red points indicate the trials from Figure 5. The third column shows the results of the responsibility condition. Here, each data point represents the average responsibility that was assigned to the black block in that trial. *Note:* 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.

model.[4] 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 that they might fall (see

[4]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.

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 relative to ground truth. 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 doesn't fall. Participants' selections were 77% accurate (67% for blocks that would fall, and 79% for blocks that wouldn't fall). For comparison, the CSM's accuracy was 79% (59%

Table 3
Correlation coefficients between individual features (or sets of features) with participants' selection judgments in Experiments 1 and 2, and their predictions in Experiment 3. Note*: The* scene features*,* black block features*,* other block features*, and* all features *rows show how well the predictions of regression models that combine these features correlate with participants' judgments. See Table 2 for a description of each feature. We fitted the features model separately on data from each experiment, or on the combined data from all three experiments (the 'All' column).*

	Experiment 1	Experiment 2	Experiment 3	All
avg_y	.16	−.01	.09	.08
avg_edge_dist	.11	−.05	−.07	.09
avg_angle	.04	.12	−.04	.13
n_blocks	−.03	−.13	.20	−.06
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	.07	.37	.05	.27
black block features	**.10**	**.39**	**.40**	**.30**
other_y	.68	.39	.57	.52
other_edge_dist	−.08	−.26	−.21	−.13
other_angle	.24	−.01	.17	.16
other_black_pile	.04	.18	-	.17
other block features	**.75**	**.56**	**.78**	**.63**
all features	**.78**	**.69**	**.84**	**.71**

would fall, 83% wouldn't fall) and the features model's accuracy was 72% (46% 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 much of the variance in participants' judgments of how many blocks would fall (Figure 6b). The features model doesn't correlate as well with participants' prediction judgments, although it is somewhat less biased (Figure 6e). While both the CSM and the features model tend to underpredict how many blocks would fall compared to participants, this bias is greater in the CSM when compared to the features model.

As Figure 7a shows, there was a tight 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 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 (out of the total number of red blocks in the scene) as a predictor for people's responsibility judgments. To map from predictions to responsibility judgments, we fit a linear regression from the proportion of blocks participants predicted to fall to their responsibility judgments. Figure 7b shows that the predictions from one group of participants are closely related to the responsibility judgments from another group ($r = .84$). This result is consistent with the idea that when evaluating how responsible the black block is, participants consider what proportion of

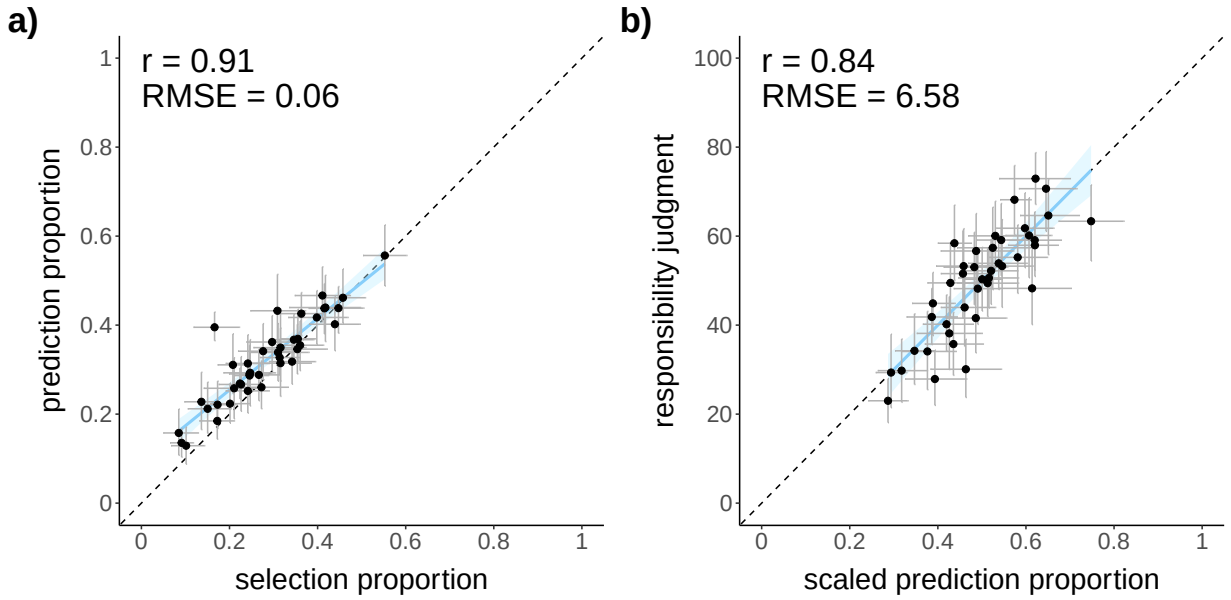

Figure 7. **Experiment 1**: Comparison between participants' responses in the three conditions. **a)** The proportion of blocks participants selected in the selection condition (x-axis) is closely related to the proportion of blocks they predicted to fall in the prediction condition (y-axis). **b)** 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 linear regression that maps from the selection proportions (which range between 0 and 1) to participants' responsibility judgments (which range 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.

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. Figures 6c and f shows how well the CSM and features model capture participants' responsibility judgments, respectively. While the CSM achieves a higher correlation than the features model, the features model has a lower error.

Table 4 shows how well individual features and sets of features correlate with participants' responsibility judgments, as well as the predictions of regressions that combine several features. The combination of scene features was a good predictor of responsibility judgments. In particular, the average y-position of each block was a strong predictor ($r = .71$).

Discussion

The results of Experiment 1 reveal a close mapping between counterfactual predictions and responsibility judgments. The greater the proportion of red blocks participants believed would fall if the black block were removed, the more responsible they judged that block to be.

Table 4
Correlation coefficients between individual features (or sets of features) with 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 red blocks in Experiments 1 and 2, and the white block in Experiment 3. See Table 2 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. The* other block features *only matter for Experiment 3 in which we asked participants how responsible the black block was for the white block staying on the table. Because Experiment 3 didn't feature any trials in which the black block was in a different pile from the white block, the* other_black_pile *feature didn't apply here. We fitted the features model separately on data from each experiment, or on the combined data from all three experiments (the 'All' column).*

	Experiment 1	Experiment 2	Experiment 3	All
avg_y	.71	.07	.09	.28
avg_edge_dist	.36	−.08	−.03	.03
avg_angle	−.15	.00	−.04	−.04
n_blocks	.13	−.03	.12	.05
scene features	**.77**	**.15**	**.17**	**.31**
black_y	−.21	−.74	−.26	−.64
black_edge_dist	−.04	.12	−.26	.07
black_angle	.05	−.04	−.31	−.03
black_above	.47	.84	.15	.69
black block features	**.52**	**.87**	**.43**	**.74**
other_y	-	-	.64	-
other_edge_dist	-	-	−.01	-
other_angle	-	-	.20	-
other_black_pile	-	-	-	-
other block features	**-**	**-**	**.75**	**-**
all features	**.84**	**.89**	**.82**	**.79**

The prediction and selection conditions directly tested participants' ability to judge what would happen if the black block were removed. Compared with the ground truth, participants in the prediction condition were less accurate (RMSE = 5.57) than participants in the selection condition (RMSE = 2.07). Having to decide whether or not each block would fall is likely to lead to a more careful consideration of what would happen than merely having to move a slider to estimate how many blocks would fall. Participants in the selection condition also took considerably more time to complete the experiment than participants in the prediction condition. Nonetheless, there was a close relationship between the proportion of blocks that participants thought would fall in both the selection and prediction conditions (Figure 7a) – and the proportion of blocks predicted to fall correlated highly with responsibility judgments (Figure 7b).

We compared participants' responses to the predictions from both the CSM and the features model (Figure 6). 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. Frequently, participants predicted that blocks above the black block would fall (see Figure 5b and d). However, there were also situations in which participants predicted that blocks that were below the black block would fall (see Figure 5a). And there were also situations in which participants anticipated longer chains of causal events, where removing the black block in one part of the tower would lead to red blocks falling off the table on the other side (see Figure 5c and e). The CSM captured participants' selections well across these situations while the features model struggled. Both models performed similarly in capturing participants' responsibility judgments. That said, a single feature was highly predictive of participants' judgments: the average y-position of the red blocks. The higher up the blocks were positioned on average, the more responsible the black block was judged to be. In Experiment 2, we constructed novel block towers to control for scene features such as the average y-position of the blocks.

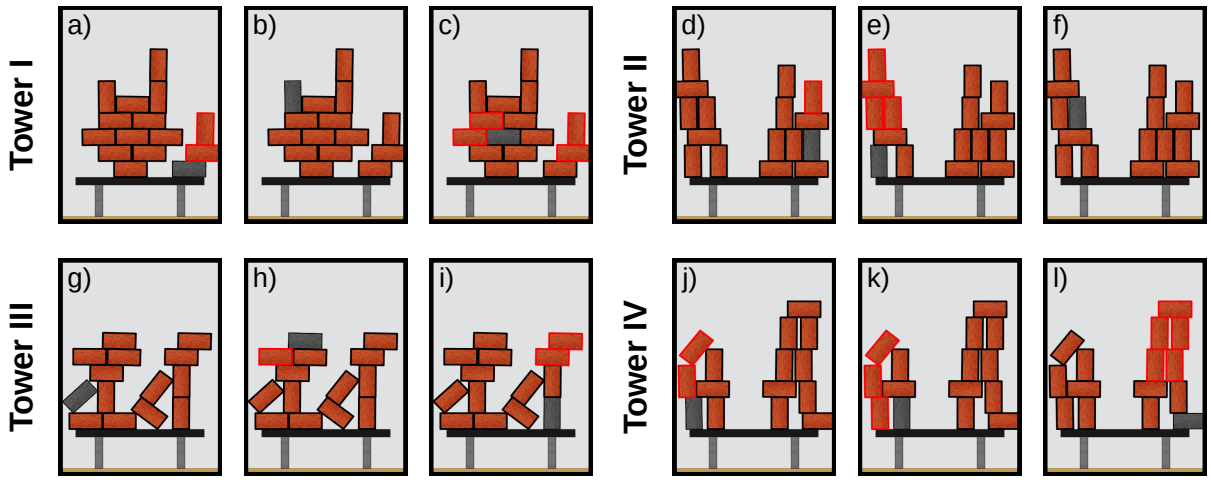

Figure 8. **Experiment 2**. Example stimuli. We created six different tower configurations (four shown), each of which was repeated seven times for different positions of the black block (three shown for each configuration). *Note:* Bright red outlines indicate which blocks would fall off the table if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

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. Because of the way in which we generated the stimuli in Experiment 1 (by dropping a pile of blocks from the top and waiting for them to settle on the table), most of the scenes featured one single pile of blocks. In Experiment 2, we used a more tightly controlled stimulus set to make sure that global scene features aren't highly correlated with the number of blocks that would fall. We also wanted to generate situations that featured piles of blocks that were disconnected from one another. We constructed 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 as these features

are identical for each tower configuration. Furthermore, while in Experiment 1 the blocks tended to form a single "pile" (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 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 ($M_{age} = 36$, $SD_{age} = 11$, 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 13.0 minutes (SD = 6.9) to complete in the selection condition, 11.6 minutes (SD = 5.2) in the prediction condition, and 7.9 minutes (SD = 3.5) in the responsibility condition.

Results

We will again discuss participants' selections, predictions, and responsibility judgments in turn.

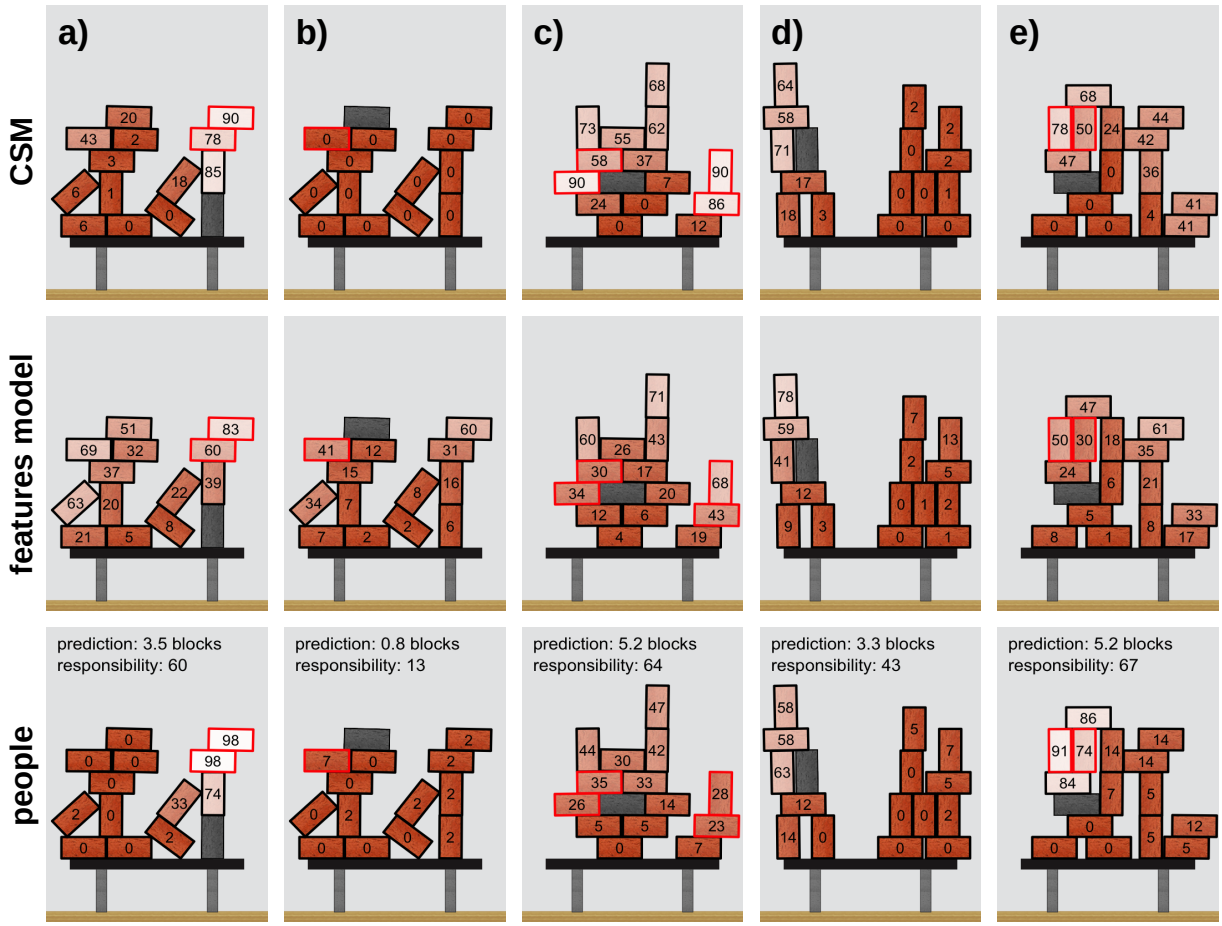

Figure 9. **Experiment 2:** Model predictions and participants' judgments for a selection of stimuli. *Note*: The number on each block indicates the percentage of participants who thought that this block would fall if the black block weren't there, and the predicted percentages for the models. The color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

Selection condition. Figure 9 shows participant responses and model predictions for a selection of stimuli. In Figure 9a and b, the CSM captures participants' selections better than the features model. In Figure 9a, the CSM assigns a high probability to the blocks above the black block falling, and a low probability to most of the other blocks. The reason why some of the blocks on the left side sometimes fall according to the CSM is because the intervention noise can lead the blocks above the black block to topple towards the left and knock against that structure of blocks. The features model overestimates the

probability that blocks would fall that participants don't select (the ones on the left in this trial), and underestimates the probability for the ones that participants do select (the ones on the right).

Figure 9b shows an example where the block tower configuration is the same as in 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 similar in Figure 9a and Figure 9b. This is because even though the black block features are changed between these scenes, the global features and the features of the red blocks are identical (see Table 2). Note that even though according to the ground truth, the block to the left underneath the black block would fall of the table, the CSM predicts that it wouldn't. This is because the CSM sets all of the objects to sleep after having applied perceptual noise. And because there aren't any blocks above the black block that would be affected by intervention noise, none of the blocks are woken up and thus stay where they are.

In Figure 9c, the features model performs better than the CSM. Participants were quite uncertain about which blocks would fall in this scene. While the features model matches participants' selections closely here, the CSM assigns a high probability to the blocks that would actually fall according to ground truth. Figure 9d shows an example in which both models perform well. Here, there are two piles of blocks disconnected from one another. Even though according to ground truth, none of the blocks would fall, both models capture participants' intuitions that some of the blocks on the left would likely fall.Figure 9e shows an example in which both the CSM and the features model perform poorly. Both models fail to match participants' high degree of certainty that the blocks above the black one would fall. Furthermore, both models assign a relatively high likelihood that the blocks on the right would fall, while participants don't think so.

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

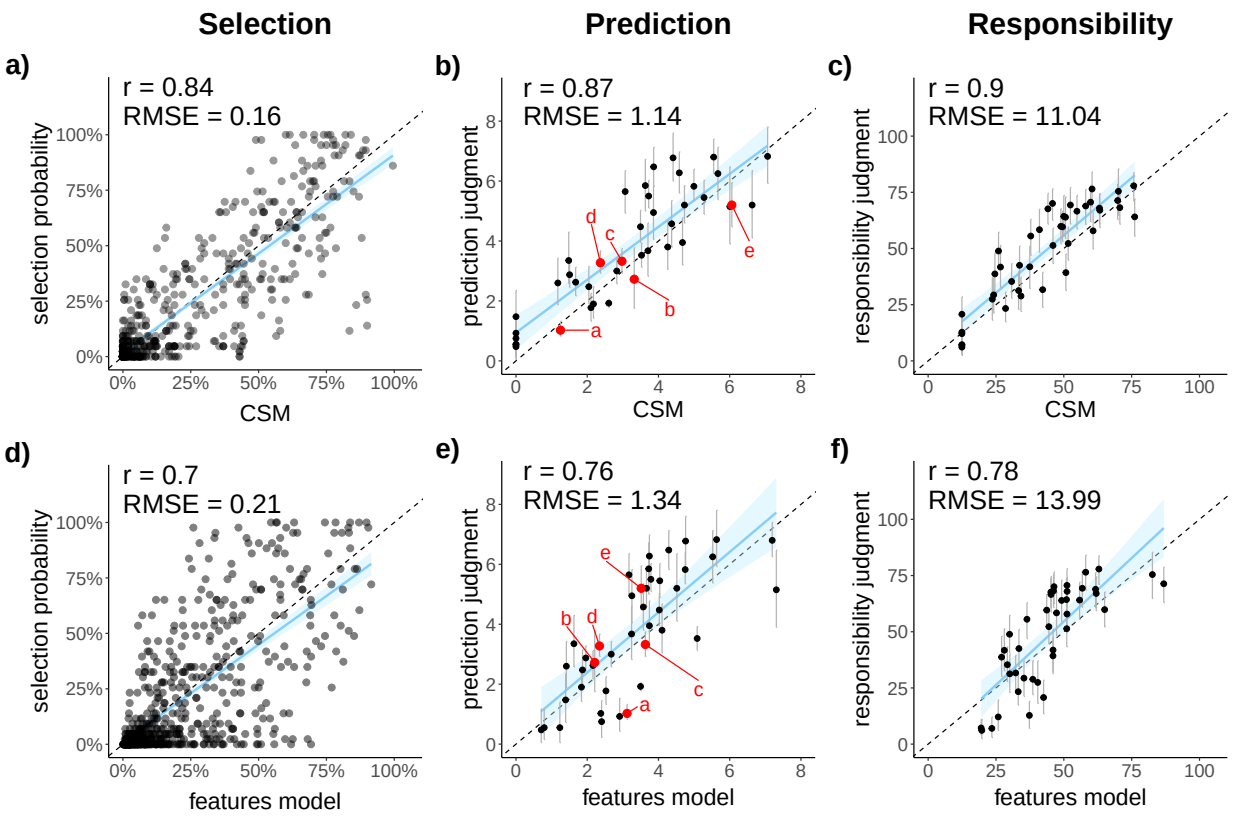

Figure 10. **Experiment 2**: Scatterplots showing the relationship between the CSM and participants' judgments at the top, and the relationship between the features model and participants' judgments at the bottom. The red points indicate the trials from Figure 9. Each point in a) and d) represents one block in one of the trials (630 blocks in total). Each point in the remaining panels represents one trial (42 trials in total). *Note:* 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.

participants' selections across all of the trials. Like in Experiment 1, the CSM captures participants' selections better than the features model. The features model again 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 83% accurate (68% for blocks that would fall, and 86% for blocks that wouldn't fall). The CSM's accuracy was 81% (63% would fall, 85% wouldn't fall), and the features model's

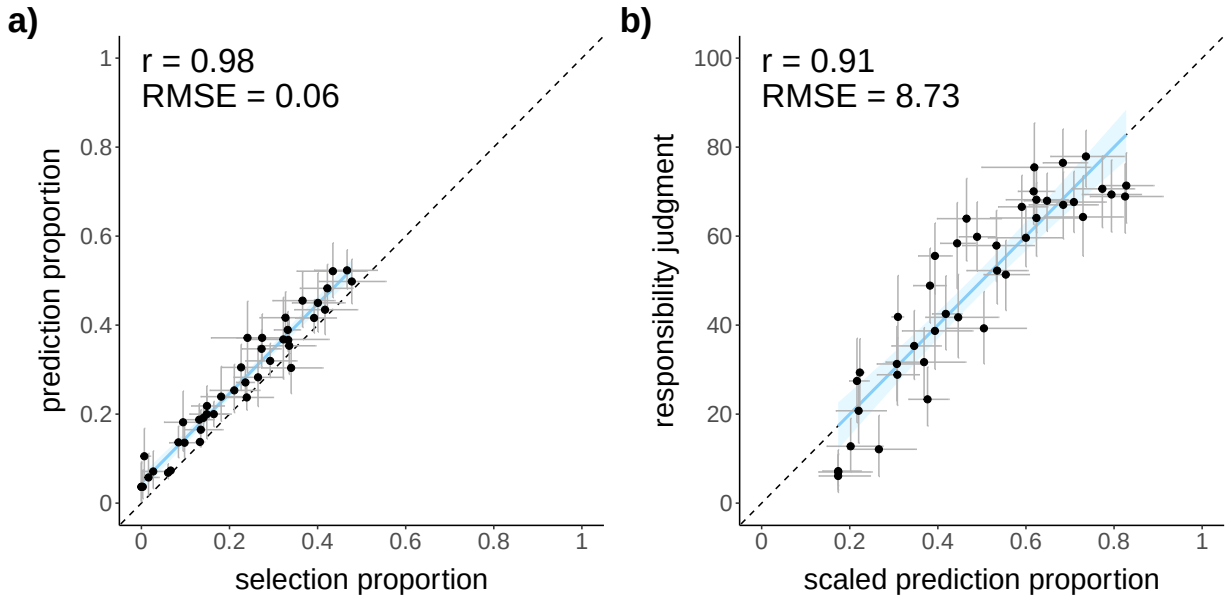

Figure 11. **Experiment 2**: Comparison between participants' responses in the three conditions. **a)** The proportion of blocks participants selected in the selection condition was compared with the proportion of blocks they judged to fall in the prediction condition. We used the proportion because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

accuracy was 76% (49% would fall, 81% wouldn't fall).

Prediction condition. Figures 10b and 10d show the relationship between model predictions and participants' predictions about how many of the red blocks would fall if the black block weren't there. The CSM does a better job of capturing participants' predictions than the features model. 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. Again, there was a very tight relationship between the number of blocks that were selected and predicted to fall, with predictions being higher than selections (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. The CSM does a better

job of capturing participants' responsibility judgments than the features model. Table 4 shows the correlations between different features and participants' responsibility judgments. As expected, 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 to be.

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. In Experiment 2, the 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 black block would have. Overall, the CSM provided a good account of participants' judgments and outperformed the features model in each of the three conditions. To gain additional insights into the role that mental simulations and features play in people's judgments, we computed a regression that combined the features model with the CSM. We applied this combined model to participants' selections and found that in Experiment 2, all of the features that were significant predictors of participants' selections become non-significant once the predictions of the CSM are added to the model (see Table A2). In this model, only the CSM is a significant predictor of participants' selections while none of the features are significant.

Even though the CSM did a good job capturing participants' judgments overall, there were some cases that reveal limitations of how the model incorporates people's uncertainty about what would happen if the black block were removed. For example, in Figure 9a, people are extremely certain that the blocks on the left side of the scene would not fall. However, the CSM predicts that some of these blocks could fall. This happens because of the way in which intervention noise is applied to the block that's removed – sometimes this noise is strong enough that the block above the black block bumps against the other blocks on the tower and thereby causes them to fall off. In Figure 9b the CSM captures participants' selections accurately. However, there are situations in which the model's predictions are likely to be wrong as it's plausible that people sometimes do in fact believe that objects underneath an object would fall if it were removed. We will return to some challenging test cases like these in the General Discussion.

Experiment 3: Judging pairs of blocks

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 ball A caused another 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 block weren't there, the more responsible the black block should be judged for the white block's staying on the table.

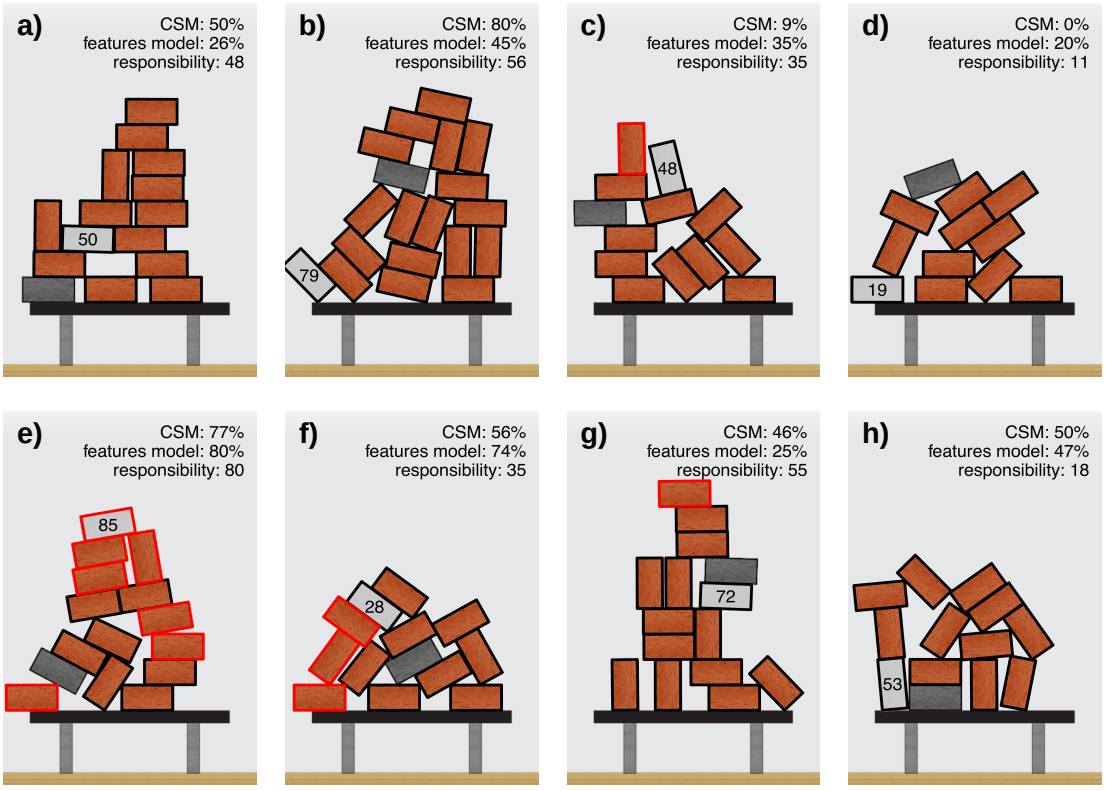

Figure 12. **Experiment 3:** Example stimuli with participants' judgments and model predictions. In the *prediction condition*, participants judged how likely the white block would be to fall if the black block weren't there. In the *responsibility condition*, participants judged how responsible the 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.

Methods

Participants. 81 participants ($M_{age} = 37$, $SD_{age} = 12$, 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 answers 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?" The sliding scale ranged 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 were 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.

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 participants 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, first focusing on some concrete cases, and then zooming out.

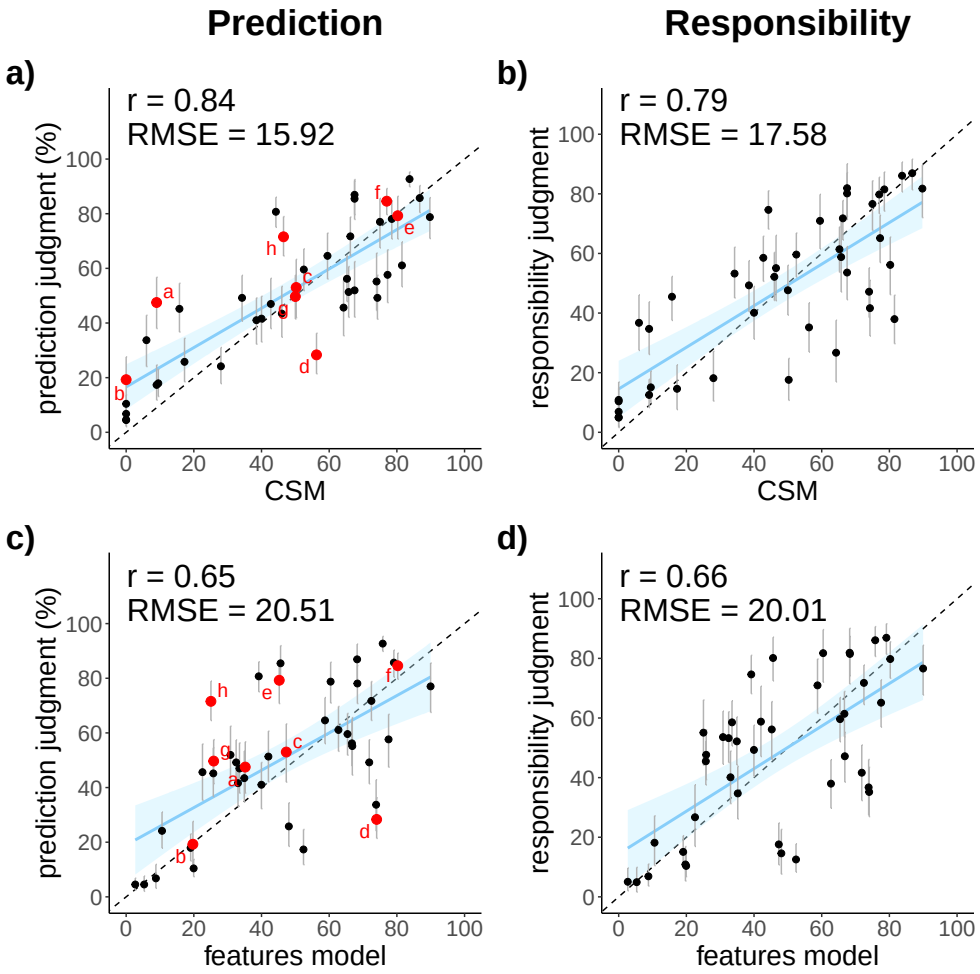

Figure 13. **Experiment 3**: Scatterplots showing the performance of the CSM and the features model on both conditions. In the prediction condition (a and c), participants were asked how likely the white block was to fall if the black block weren't there. In the responsibility condition (b and d), participants were asked how responsible the black block was for the white block staying on the table. Red letters indicate the trials from Figure 12. *Note:* Error bars indicate bootstrapped 95% confidence intervals. The blue ribbon shows the 95% confidence interval of the regression line.

Prediction condition. Figures 12a and b show two cases in which the predictions of the CSM match participants' judgments better than those of the features model. The features model assigns a low probability because in general, the white block is less likely to fall when its vertical position is low, or when the black block's position is high. In Figures 12c and d the features model matches participants' judgments more closely than the CSM. In Figure 12d the CSM assigns a probability of 0% that the white block would

fall. Here, there aren't any red blocks above the black block that could be affected by the intervention noise. This means that only perceptual noise is applied but because the objects are put to sleep after that happens, and no collisions take place to wake the objects back up, none of the blocks fall off the table (similar to CSM's judgments in Figure 9b). Figure 12e and f show situations in which both models capture participants' judgments well, and Figure 12g and h show situations in which neither model performs well. In Figure 12g, the white block is directly *under* the black block, and both models underestimate people's judgments in this case. In Figure 12h both models capture participants' predictions well, but there is a mismatch between prediction and responsibility judgments.

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 extreme than what the models predict (as indicated by the regression line being off the diagonal). To get a sense of 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. 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. Table 4 shows which features best correlated with participants' responsibility judgments. This time, the best predictor is the y-position of the white block. The higher that block was positioned, the more responsibility participants tended to assign to the black block.

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

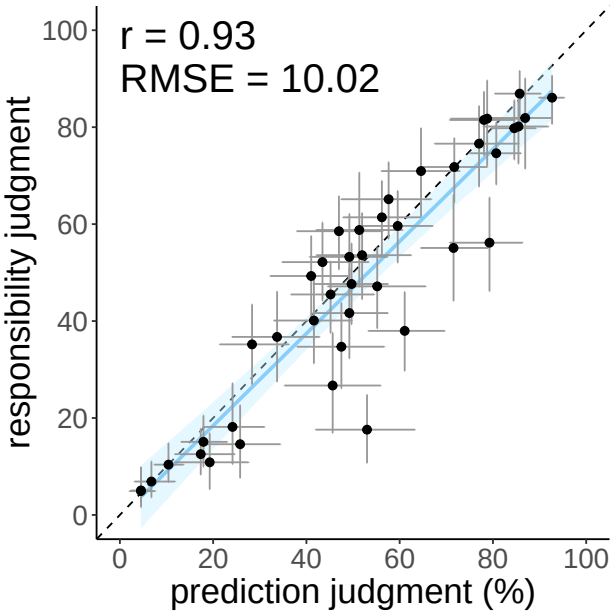

Figure 14. **Experiment 3:** Relationship between participants' predictions of how likely the white block would be to fall if the black block were removed (x-axis) and the extent to which the black block was judged to be responsible for the white block staying on the table (y-axis). Each point shows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

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 to be.

Discussion

While Experiments 1 and 2 investigated 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 black 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 very closely related ($r = .93$). 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 in 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 may have 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 the diffusion of responsibility between agents).

In Experiments 1 and 2, the responsibility question was somewhat ambiguous. We had asked participants 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 however possible that some participants interpreted the question differently and, for example, cared about the absolute number instead. 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 some candidate event (or object) caused a particular event to happen. For example, the question might be whether billiard ball A caused 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 asked 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 judging responsibility for stability and considering counterfactual simulations are intimately related. The features model which predicts participants' judgments without relying on physical simulations didn't fare as well. In Experiment 1, scene features, such as the average distance of each 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 are 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 while there are features in each experiment that are associated with participants' judgments, how much each feature matters changes between experiments. In contrast, the CSM provides a unified account of participants' judgments across all experiments.

While the CSM captures much of the variance in participants' responsibility judgments in Experiment 3, it did so slightly less well than in Experiments 1 and 2. This may seem somewhat surprising given that Experiment 3 was most closely modeled after the situations in which the CSM was originally developed, namely where the question is to what extent a single candidate cause was responsible for an event of interest. Note, however, that in Experiments 1 and 2, there is some room for the model to get things right for the wrong reasons, and that's not possible in Experiment 3. In Experiments 1 and 2, the CSM uses the proportion of blocks predicted to fall to determine responsibility. It's

possible that it sometimes gets the proportion right but actually predicts that different blocks would fall than the ones that people think would fall. In Experiment 3, rather than predicting the proportion of blocks that would fall, the CSM has to predict the probability with which a single block would fall, and there is less room to get that right for the wrong reasons.

Most importantly, while there is still some room for improving how the CSM captures participants' simulations of what would happen, there was a very close relationship between participants' predictions and responsibility judgments (see Figure 14). And this really is the main claim of the CSM: people judge responsibility by considering what would have happened in a relevant counterfactual situation.

Model comparison

The full CSM includes three sources of uncertainty that affect people's predictions about what would happen. To assess 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. 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 evaluated the models' performance on the selection data from Experiments 1 and 2 and the prediction data from Experiment 3. Table 5 shows the results of this analysis.

The full CSM outperforms all of the lesioned models: it correlates more strongly with participants' responses and has a lower error in the held-out test sets. In a sensitivity analysis in the appendix (see Figure A1), we show how the CSM's performance depends on

the parameter values. The loss landscape is smooth in that small changes to the parameters don't lead to big changes in the model fits (as desired). 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 outperforms the features model. The features model achieves a high correlation with participants' judgments when fitted separately to the different experiments (see Table 3). However, there is no single parameter setting that works well across all three of the experiments. Which features matter most differs between the experiments.

Table 5

Model comparison results on aggregated data from the selection condition of Experiments 1 & 2 and the prediction condition of Experiment 3. The model column specifies the model version. The best-fit column shows the best-fitting parameter values for each model (when the model is fitted on all of the data). An interactive widget for visualizing different versions of the CSM is available at `https://cicl-stanford.github.io/mental_jenga/interface`*. The cross-validation results show the median and the 5% and 95% quantiles for each test set across the 200 cross-validation runs. The r column shows the correlation between model predictions and participants' responses. The Δr column shows the difference in r between the full CSM and the other models. The RMSE column shows the root mean squared error between model prediction and participants' responses. The $\Delta RMSE$ column shows the difference in RMSE between the full CSM and other models. The AIC and BIC columns show the Akaike Information Criterion and the Bayesian Information Criterion for each model when fitted on all of the data. For these two measures, lower values indicate better model performance.*

		split-half cross-validation					
model	best-fit	r	Δr	RMSE	ΔRMSE	AIC	BIC
CSM (p, i, d)	2, 5, 4	.86 [.84, .87]	-	2.53 [2.25, 2.83]	-	938	953
CSM (p, i)	2, 5	.84 [.83, .86]	.01 [.00, .02]	2.72 [2.44, 3.06]	0.19 [0.01, 0.43]	949	959
CSM (p, d)	4, 3	.76 [.73, .79]	.10 [.07, .12]	4.35 [4.03, 4.69]	1.82 [1.41, 2.19]	1105	1115
CSM (i, d)	5, 5	.85 [.83, .86]	.01 [.00, .02]	2.77 [2.45, 3.13]	0.24 [0.06, 0.43]	989	1004
CSM (p)	5	.74 [.72, .77]	.11 [.08, .13]	4.45 [4.08, 4.81]	1.92 [1.50, 2.31]	1094	1099
CSM (i)	5	.84 [.82, .85]	.02 [.01, .03]	2.90 [2.58, 3.27]	0.38 [0.17, 0.58]	1003	1008
CSM (d)	9	.68 [.65, .71]	.17 [.14, .20]	6.79 [6.21, 7.40]	4.27 [3.67, 4.88]	1543	1548
features	-	.72 [.79, .74]	.14 [.11, .17]	4.37 [3.97, 4.83]	1.84 [1.31, 2.32]	1036	1101

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

Responsibility judgments

We performed separate cross-validations to evaluate whether the CSM also outperforms the features model in capturing participants' responsibility judgments. For this analysis, we compared the full CSM with the features model. Because the responsibility question differed between some of the experiments, we ran two separate cross-validation analyses: one that combined the data from Experiments 1 and 2, and one for Experiment 3.

For the cross-validation on Experiments 1 and 2, we determined the best-fitting version of the CSM on each training set by first calculating the proportion of blocks it predicted to fall on each trial in the training set for each parameter setting of the CSM (defining a grid over the noise parameters β_p, β_i, and β_d). We then computed a linear regression, mapping from the proportion of blocks predicted to fall to participants' responsibility judgments. This procedure fits five parameters in total: the three noise parameters in the CSM, and the intercept and slope in the linear regression. We find the five parameters that minimize the squared loss between model predictions and responsibility judgments on the training set, and then, using these parameters, compute the squared loss on the held-out test set. For the features model, we fit a linear regression using nine parameters on the training set (four scene features, four black block features, and one intercept; see Table 2) and then compute the loss on the test set. We repeat this procedure 200 times using split-half cross-validation. Table 6 shows the results of this analysis. The CSM captures participants' responsibility judgments better than the features model as indicated by positive values of Δr and $\Delta RMSE$.

We performed a separate cross-validation for Experiment 3. In this experiment, the CSM predicts that the black block's responsibility is a direct function of how likely the white block would fall. So for this experiment, the CSM only has its three noise parameters, and no mapping via a linear regression is required. The features model has twelve parameters in this experiment because it also encodes information about the white

block (see Table 2). As Table 6 shows, the CSM captures participants' responsibility judgments better than the features model.

Error analysis of prediction judgments

As an additional test for how well the CSM and the features model capture participants' judgments, we can compare the errors that people make with those of the models. For this analysis, we focused on the prediction conditions in Experiments 1 and 2. We computed the model error by subtracting the model predictions from ground truth. For example, if, according to ground truth, 5 red blocks would fall if the black block were removed, but a model predicts that only 3 blocks would fall, this would be a prediction error of -2.[5] Figure 15 shows scatter plots between the prediction errors from models and people. The prediction errors between models and people are strongly correlated in both experiments. In both experiments, the prediction errors between CSM and people are more strongly correlated with one another than the errors between the features model and people. However, in Experiment 1, the features model is less biased (lower RMSE) than

Table 6
Model comparison results for the full CSM and features model when fit directly to the responsibility condition. The cross-validation results show the median and the 5% and 95% quantiles for each test set across 200 cross-validation runs. The r column shows the correlation between model predictions and participants' responses. The Δr column shows the difference in r between the CSM and the features model. The RMSE column shows the root mean squared error between model prediction and participants' responses. The $\Delta RMSE$ column shows the difference in RMSE between the CSM and the features model.

Exp	Model	r	Δr	RMSE	ΔRMSE
1 & 2	CSM	.87 [.83, .91]	-	9.37 [8.02, 10.57]	-
	features	.78 [.67, .85]	.08 [.00, .16]	11.52 [9.50, 13.55]	2.15 [−0.26, 4.81]
3	CSM	.85 [.78, .90]	-	18.12 [13.72, 21.61]	-
	features	.59 [.36, .77]	.19 [.01, .41]	25.02 [18.79, 35.46]	6.90 [−1.15, 16.76]

[5]This kind of analysis wouldn't be very informative for Experiment 3 as the ground truth is just 1 or 0 depending on whether the white block falls, and so there is no way for a model to make an error in both directions from the ground truth.

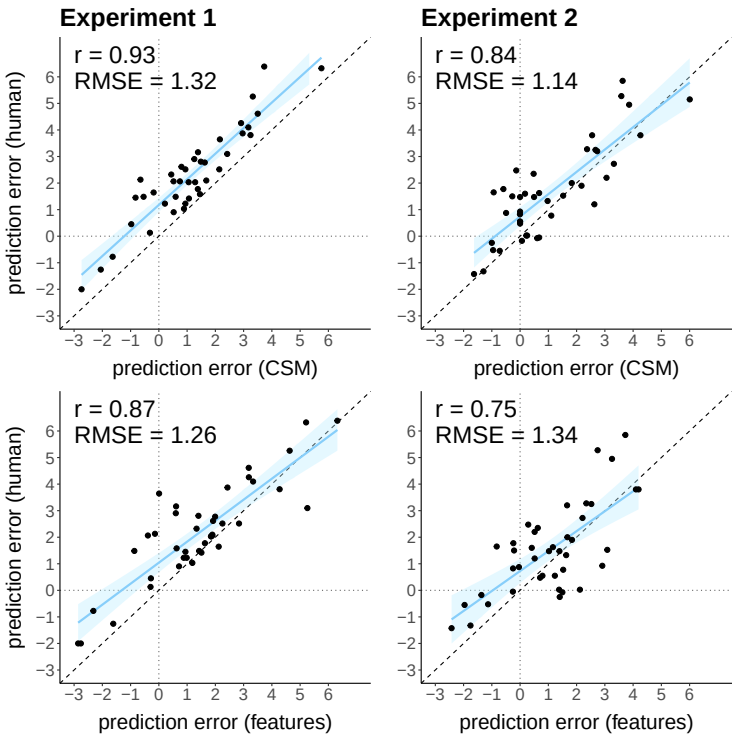

Figure 15. **Error analysis:** Relationship between participants' errors in the prediction condition, and both CSM and features model errors when fit to selection condition data, for Experiments 1 and 2. Each point shows the averaged judgments for one trial, and the error bars are 95% bootstrapped confidence intervals.

the CSM. As we already saw in Figure 6b and e, the CSM underestimates more strongly than the features model how many blocks people think would fall.

Individual participant analysis

We used the CSM to model participants' aggregated judgments. For example, in the selection conditions of Experiments 1 and 2, the model predicts the probability that a particular block would be selected by participants. For any given noise parameter setting, the model approximates this "true" probability by running a large number of simulations. We chose 200 simulations for practical reasons. A larger number of simulations would simply result in a better estimate of that probability. Of course, we do not assume that this is a model of how individual participants select which blocks would fall. It's much more plausible that participants only run a very small number of simulations. Gerstenberg et al.

(2017) used eye-tracking to quantify how many counterfactual simulations participants produced when assessing whether ball A caused ball B to go through a gate. Generally, the number was small (no more than 3), and greater in situations in which the counterfactual outcome was more uncertain (see also Hamrick, Smith, Griffiths, & Vul, 2015).

How many simulations did participants run in our experiments? While we cannot know for sure, we can at least approximate an answer. To do so, we used a procedure that Battaglia et al. (2013) developed: we simulated artificial participants from our model and varied how many simulations (N) each individual participant generated before making a judgment. For example, if a participant ran two simulations in the selection condition, any given brick would either fall zero, one, or two times. The participant then chooses to select blocks based on these simulations (here basically flipping a coin for the blocks that fell once). We can then compare the aggregated responses from these simulated participants who run N simulations before making a judgment to those of our actual participants. Based on this analysis we found that our results are most consistent with individual participants running a single simulation to make their judgment (Vul, Goodman, Griffiths, & Tenenbaum, 2014; see Figure A2).

General Discussion

How do people judge whether one object is responsible for the stability of another? In this paper, we developed the counterfactual simulation model (CSM) of causal judgments about physical support. The CSM predicts that people judge physical support by mentally simulating what would happen if the object of interest were removed. We tested the CSM across three experiments in which participants made judgments about towers of blocks stacked on a table. Similar to how people spontaneously consider counterfactuals when judging causation for dynamic physical events (Gerstenberg et al., 2017), the CSM assumes that people play "Jenga in their mind" when judging responsibility in static scenes involving 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 its stability.

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, in press; 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 a 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 compares 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. 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, 2022, 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" (see Table 1). 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

et al., 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 role of mental simulation in assessing physical support

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., in press; 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 when evaluating the extent to which the black block is responsible. 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.

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, as 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 tabletop 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 of stability in many situations.

While the current version of the CSM relies heavily on the process of simulation, it also encodes more general heuristic principles in the way in which it applies uncertainty. Specifically, it assumes that counterfactual interventions initially only affect the objects that are above the black block, and it uses a heuristic definition of what 'above' means. Furthermore, it uses the heuristic principle of putting objects to sleep so as to only simulate those parts of the scene that would be affected by the black block's removal (Ullman et al., 2017). Future work is required to delineate more clearly how visual features, heuristic principles, and mental simulations jointly contribute to participants'

physical predictions. In our experiments, participants had as much time as they liked to respond. It would be interesting to explore what participants' judgments would look like under time pressure, or if they are put under cognitive load with a secondary task. It's possible that in such settings, people would rely more heavily on general features of the scene, and less so on the process of mental simulation.

A simulation model is a powerful tool. We focused here on how simulations enable judgments of physical support. 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 would happen next. A flexible simulation engine provides a unified framework for modeling a great variety of different tasks. For example, participants could be asked to add a block to make a tower more robust or to remove a block to make the greatest number of blocks fall off the table. We could also ask participants to choose a block to remove such that exactly N other blocks would fall off the table, that a specific set of blocks would fall off, or that one specific block would fall off while another specific block would stay on, etc.

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 often go together, they can also come apart.

Figure 16a and b show two examples in which the black block prevents the white block from falling off the table. In both situations, the white block would fall off the table

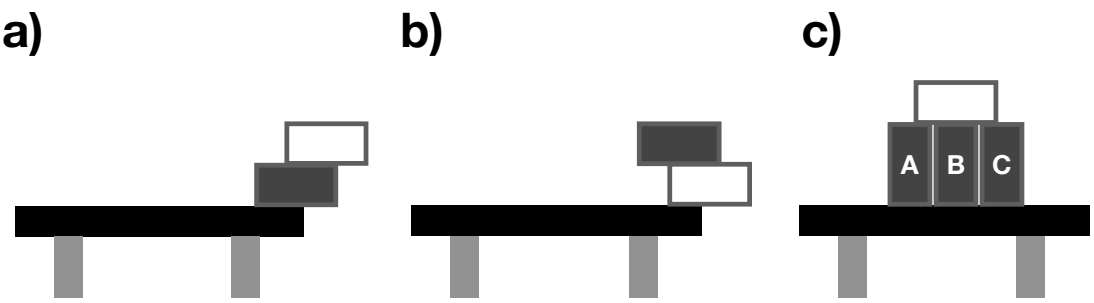

Figure 16. Scenes illustrating interesting border cases of physical support. In a) and b), the black block is responsible for the white block staying on the table. The white block would fall off the table if the black block weren't there. However, only in a) but not in b) does it seem appropriate to 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.

if the black block weren't there. However, while it feels right in Figure 16a to say that the black block supports the white block, it doesn't feel right to say so in Figure 16b. For the black block to support the white block, it must be positioned underneath.

There is another sense in which supporting and preventing from falling come apart. Intuitively, all that's required for object A to support object B is that object B would move (ever so slightly) if object A were removed. Accordingly, in many of our stimuli, the black block supports many more objects than it prevents from falling off the table. The CSM could be adapted to test for physical support in the following way: first, consider a counterfactual intervention that removes the object of interest from the scene (with some small perturbation), then check for every object (above the removed one) whether its position changed from what it was before the intervention. Determining physical support in this way would arguably be less challenging than assessing whether a block prevented another from falling off the table. To assess support, one only needs to simulate a few steps ahead, whereas for assessing prevention from falling, one needs to run the simulation many more steps. Battaglia et al. (2013) showed that participants' have an easier time judging whether or not a tower will fall (and in which direction it will fall) than judging how far

the blocks will fall. Whereas judging whether a block would fall (and in what direction) only requires simulating a few steps forward, judging how far the blocks will fall requires simulating many more steps.

The case shown in Figure 16b also highlights a potential limitation of the way in which the CSM is implemented. Because of the way in which uncertainty is implemented in the CSM, it actually wouldn't predict that the white block would fall. Remember that the model applies perceptual noise first, puts the objects to sleep, adds intervention noise by applying an impulse to the blocks above the black one, and then uses dynamic noise to resolve the collisions that unfold. In this case here, there are no blocks above the black block. What this means is that while perceptual noise is applied to the white block, it's put to sleep afterward and thus doesn't fall off the table (because there aren't any other objects whose collisions could wake it up). Only objects that are awake move.

Some of these situations arose in our experiments. For example, in the tower in Figure 9b of Experiment 2, the black block is at the very top. According to the ground truth model, the left block underneath it would fall off the table. However, 3 out of 43 participants believed that this block would fall. Most participants thought that none of the blocks would fall off. Similarly, in the tower in Figure 12d of Experiment 3, only a few participants thought that the white block on the bottom left would fall if the black block were removed. In both of these cases, the CSM is certain that none of the blocks would fall (and thus captures people's judgments quite well).

Of course, one could consider alternative ways of implementing sleep in the model. For example, sleep could be implemented in a way such that not just the objects that are above the black block are woken up, but all the objects that are in contact with it. Implemented in this way, the CSM would predict that the blocks in the two examples we discussed would fall. It would thus be closer to ground truth, but further away from people's intuitions. People's tendency to mostly consider what consequences the removal of an object would have to the objects above, and less so to the objects below, is a bias in

physical reasoning worthy of further exploration.

Overdetermination

One of the examples of physical support from the Oxford Dictionary that we mentioned earlier states that "the dome was supported by a hundred white columns". It's plausible that the dome would not collapse if one of the columns were removed. However, we would still want to say that each of the columns supports the dome. 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 16c. 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, if any of the three blocks individually were removed, the white block would not fall.

In the literature on causation, a scenario like the one depicted in Figure 16c 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 hadn't been there, then block A would have been pivotal.

The current version of the CSM assigns 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 when a black block is removed. The model predicts that a block's responsibility will reduce when other blocks are present in the scene that could stop the target block from falling. 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).

Situations of overdetermination also arise in dynamic physical interactions. For example, imagine that one ball C lies in front of the gate, and then two balls, A and B, collide with ball C at the same time and knock it into the gate. The situation is such that either ball A or ball B would have been individually sufficient to knock ball C into the gate. Did ball A cause ball C to go into the gate? Intuitively, the answer is 'yes' even though ball C would still have gone through the gate without ball A. In Gerstenberg, Goodman, et al. (2021) we handle these kinds of cases by assuming that people's causal judgments are

sensitive to different aspects of causation that map onto different counterfactual tests. For example, one counterfactual test removes the candidate object from the scene and simulates what would have happened without it. We call this aspect WHETHER-CAUSATION as it reveals whether the candidate cause made a difference to whether or not the outcome. Another counterfactual test slightly perturbs the candidate object and simulates whether the outcome event (finely construed) would have been different from what actually happened. We call this aspect HOW-CAUSATION as it reveals whether the candidate cause made a difference to how the outcome came about. Finally, we consider a counterfactual test for whether the cause object was sufficient for bringing about the outcome. To test for SUFFICIENT-CAUSATION, the CSM considers a counterfactual situation in which alternative causes were removed from the scene and checks whether the object of interest would have caused the outcome in that situation. In the overdetermination scenario, ball A is not a WHETHER-CAUSE of the outcome (because ball B would have knocked ball C into the gate without ball A). However, ball A is a HOW-CAUSE because the outcome event would have been slightly different had ball A been perturbed (e.g., ball C would have gone through the gate at a slightly different location and at a slightly different point in time). Ball A was also a SUFFICIENT-CAUSE of the outcome. In a situation in which the alternative cause (ball B) had been removed, ball A would have caused ball C to go into the gate.

Gerstenberg, Goodman, et al. (2021) demonstrated how the different aspects of causation help explain participants' judgments across a challenging set of test cases. Only considering WHETHER-CAUSATION by itself was not sufficient to account for participants' causal intuitions. In this paper, our implementation of the CSM draws on some of these aspects of causation. The way in which we model people's uncertainty about the intervention includes both the removal of the black block as well as a perturbation to the blocks above it. Modeling intervention uncertainty in this way is reminiscent of WHETHER-CAUSATION and HOW-CAUSATION. However, currently, the model doesn't test for whether the black block was sufficient to guarantee some other objects' stability. To

adequately deal with situations of causal overdetermination, such as the example shown in Figure 16c, incorporating a test for sufficiency may be required.

Conclusion

Humans have a remarkable grasp of 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 dynamic physical events (Gerstenberg, Goodman, et al., 2021), omissions (Gerstenberg & Stephan, 2021), and static scenes involving physical support. 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 (Sosa, Ullman, Tenenbaum, Gershman, & Gerstenberg, 2021; Wu, Sridhar, & Gerstenberg, 2022). In this paper, we investigated people's judgments about block towers as a case study. However, physical support manifests itself in all sorts of ways: from blocks supporting towers to rocks supporting houses to socks supporting legs. More work is needed to explore how well the CSM generalizes to other domains and tasks.

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

Ahuja, A., & Sheinberg, D. L. (2019). Behavioral and oculomotor evidence for visual simulation of object movement. *Journal of Vision*, *19*(6), 13–13.

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

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.*

Beller, A., Xu, Y., Linderman, S., & Gerstenberg, T. (2022). Looking into the past:

Eye-tracking mental simulation in physical inference. *Cognitive Science Proceedings*.

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.

Crespi, S., Robino, C., Silva, O., & de'Sperati, C. (2012). Spotting expertise in the eyes: Billiards knowledge as revealed by gaze shifts in a dynamic visual prediction task. *Journal of Vision*, *12*(11), 1–19.

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

Gerstenberg, T. (in press). What would have happened? counterfactuals, hypotheticals, and causal judgments. *Philosophical Transactions of the Royal Society B: Biological Sciences*. 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.

Hamrick, J. B., Smith, K. A., Griffiths, T. L., & Vul, E. (2015). Think again? the amount of mental simulation tracks uncertainty in the outcome. In D. C. Noelle et al. (Eds.), *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (pp. 866–871). Austin, TX: Cognitive Science Society.

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

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).

Kahneman, D., & Tversky, A. (1982). The simulation heuristic. In D. Kahneman & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 201–208). New York: Cambridge University Press.

Kominsky, J. F., & Phillips, J. (2019, Oct). Immoral professors and malfunctioning tools: Counterfactual relevance accounts explain the effect of norm violations on causal selection. *Cognitive Science*, *43*(11), e12792. Retrieved from `http://dx.doi.org/10.1111/cogs.12792` doi: 10.1111/cogs.12792

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. (2020). Broken physics: A conjunction-fallacy effect in intuitive physical reasoning. *Psychological Science*, *31*(12), 1602–1611.

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: Naíve 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.

Phillips, J., & Knobe, J. (2018). The psychological representation of modality. *Mind & Language*, *33*(1), 65–94.

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. F. (2022). Irreversible (one-hit) and reversible (sustaining) causation. *Philosophy of Science*, *89*(5), 889–898.

Rule, J. S., Tenenbaum, J. B., & Piantadosi, S. T. (2020). The child as hacker. *Trends in cognitive sciences*, *24*(11), 900–915.

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. (in press). 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.

Sosa, F. A., Ullman, T. D., Tenenbaum, J. B., Gershman, S. J., & Gerstenberg, T. (2021). Moral dynamics: Grounding moral judgment in intuitive physics and intuitive psychology. *Cognition*. Retrieved from `https://psyarxiv.com/xh4kg`

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

Vul, E., Goodman, N., Griffiths, T. L., & Tenenbaum, J. B. (2014, Jan). One and done? optimal decisions from very few samples. *Cogn Sci*, *38*(4), 599-637. Retrieved from http://dx.doi.org/10.1111/cogs.12101 doi: 10.1111/cogs.12101

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).

Wu, S., Sridhar, S., & Gerstenberg, T. (2022). That was close! a counterfactual simulation model of causal judgments about decisions. *Cognitive Science Proceedings*.

Yi, K., Gan, C., Li, Y., Kohli, P., Wu, J., Torralba, A., & Tenenbaum, J. B. (2019). *Clevrer: Collision events for video representation and reasoning*.

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

Pairwise feature correlations

Table A1

Correlation coefficients between individual features across all trials in all of the experiments. See Table 2 for a description of each feature. Note*: We abbreviated the feature labels in this table here. One of our criteria for including features in the model was that none of the pairwise features correlations was greater than* $r = 0.8$*. In fact, the strongest correlation was between the number of blocks above the black block, and the y-position of the black block with* $r = -.61$*: the higher the black block was in a tower, the fewer blocks were on top of it.*

		scene features				**black block features**				**other block features**			
		y	**edge**	**angle**	**blocks**	**y**	**edge**	**angle**	**above**	**y**	**edge**	**angle**	**pile**
scene	**y**	-											
	edge	−.21	-										
	angle	−.43	.15	-									
	blocks	.48	−.26	−.42	-								
black	**y**	.06	−.09	−.01	.03	-							
	edge	−.14	.58	.11	−.17	.10	-						
	angle	−.13	.12	.43	−.22	−.05	.06	-					
	above	.21	.20	−.10	.18	−.61	.25	−.15	-				
other	**y**	.20	−.03	−.08	.09	−.05	−.05	−.03	.08	-			
	edge	−.10	.47	.06	−.13	−.05	.23	.05	.09	.13	-		
	angle	−.21	.08	.49	−.20	−.00	.06	.16	−.03	.05	.10	-	
	pile	−.19	.55	.15	−.05	−.05	.30	.07	.09	−.01	.36	.11	-

Sensitivity analysis

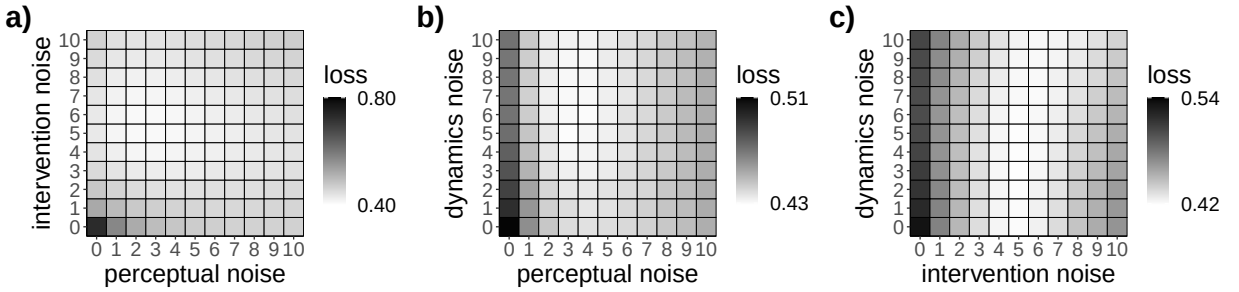

Figure A1. Sensitivity analysis of the noise parameters in the full CSM. Heatmaps showing averaged cross-sections of our 3-way grid search over the parameter space for the full CSM. Note that the range for the gradient is different in each of the panels. The model was evaluated on the selections from Experiments 1 and 2, and the predictions from Experiment 3. We performed a grid search over the perceptual noise, intervention noise, and dynamic noise, $\{\beta_p, \beta_i, \beta_d\} \in [1, 2, ..., 10]^3$, totaling 1331 sets of parameters. For each parameter setting, we ran 200 simulations in Experiments 1 and 2, and 400 simulations in Experiment 3, for each trial in those experiments. In **a)** the losses are averaged over all β_d, in **b)** the losses are averaged over all β_i, and in **c)** the losses are averaged over all β_p. The best-fitting CSM has the following parameter values: $\beta_p = 2, \beta_i = 5, \beta_d = 4$. An interactive widget for visualizing different versions of the CSM is available at `https://cicl-stanford.github.io/mental_jenga/interface`. Overall, the results show that the different sources of noise partly trade off. For example, a) shows that at least some intervention noise or some perceptual noise is required to improve the model's performance. Dynamic noise affects the model's performance less strongly than the other sources of noise. For example, in b) and c) the loss gradient is stronger along the x-axis versus the y-axis. This result is in line with the cross-validation results in Table 5, showing that models which drop the dynamic noise generally fare better than models that drop either of the other two sources of noise.

CSM and features model combined

Table A2

*Results of features model fits comparing the features model and a "hybrid" model that includes the features together with the predictions of the full CSM (features + CSM). The models were fitted to participants' selections in Experiment 1 and Experiment 2. The rows above the double line at the bottom, show the coefficients for each (normalized) predictor. Bolded values indicate significant predictors with $p < 0.01$. The r values below the double line indicate how well a model fitted on one (or all experiments) correlates with participants' judgments in other experiments. For example, the "**r** (Experiment 1)" row shows how well a model that was fit to Experiment 1 correlates with participants' judgments in Experiment 2 ($r = .57$ for the features model, and $r = .83$ for the features + CSM). Overall, these analyses show that once the CSM is added as a predictor to the model, most of the other features become non-significant predictors of participants' judgments.*

	features			**CSM**	**features + CSM**		
	Exp 1	Exp 2	All	All	Exp 1	Exp 2	All
`intercept`	**−1.24**	**−1.77**	**−1.50**	-	**−2.10**	**−2.98**	**−2.49**
`avg_y`	.17	.58	.16	-	.16	.35	.13
`avg_edge_dist`	.19	−.09	.16	-	.09	−.35	0
`avg_angle`	.01	.13	.20	-	.10	.26	.17
`n_blocks`	−.32	−.85	**−.27**	-	−.19	−.37	−.17
`black_y`	−.21	**−.46**	**−.38**	-	−.18	−.26	−.26
`black_edge_dist`	−.05	**.47**	−.26	-	−.02	.30	.15
`black_angle`	0	.05	−.01	-	.02	.29	.05
`black_above`	.03	.23	.24	-	−.23	−.02	−.05
`other_y`	**1.45**	**.83**	**1.12**	-	**.64**	.22	**.39**
`other_edge_dist`	**−1.01**	**−.91**	**−.87**	-	**−.51**	−.10	−.27
`other_angle`	.29	−.09	.17	-	.09	.06	.07
`other_black_pile`	.45	**.62**	**.58**	-	.21	.22	.26
`CSM`	-	-	-	-	**3.68**	**4.72**	**4.06**
r (Experiment 1)	.78	.57	.76	.87	.88	.83	.87
r (Experiment 2)	.59	.73	.70	.84	.81	.86	.86
r (Experiment 3)	.77	.40	.65	.80	.83	.74	.80
r (All)	.67	.65	.73	.86	.84	.84	.86

Number of simulations

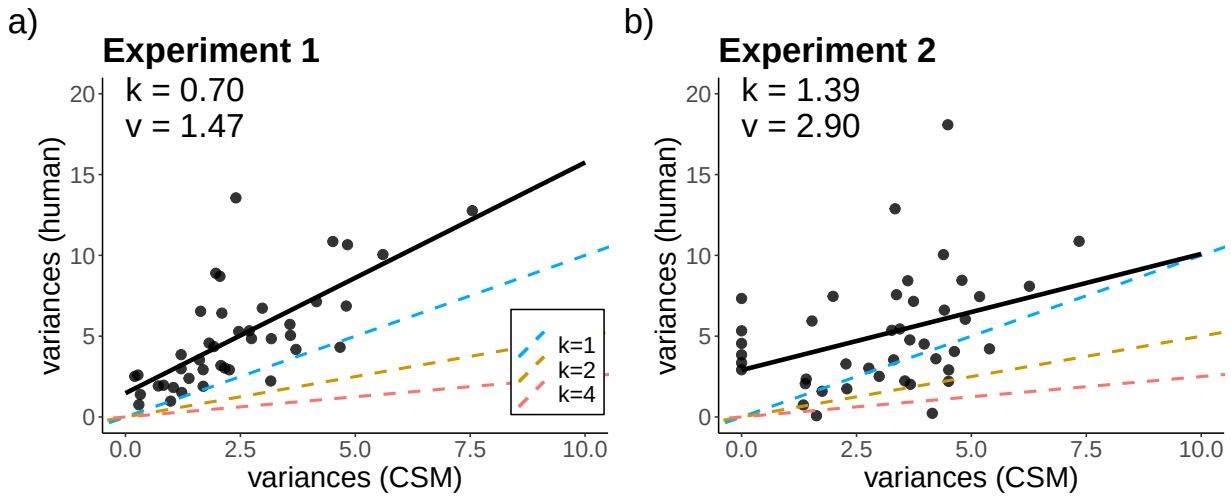

Figure A2. Relationship between the variance in participants' responses in the selection condition and the CSM's corresponding judgment as a function of how many simulations k the CSM assumes each participant ran. We assume that participants run k simulations and make their predictions by averaging the number of blocks that fall across these simulations. Here, we estimate k by noticing that the variance in participants' responses is a function of k: running more simulations should correspond to reduced variance. Variance in participants' judgments, however, may stem from multiple sources. In line with Battaglia et al. (2013), we model the variance in participants' judgments as arising from two sources: *sampling variance* (that we are interested in), and a general *judgment variance*. The sampling variance for running k simulations on any particular trial is equal to $(1/k)v_0$, where v_0 is the sampling variance of a single simulation and is unique in each trial. In our analysis, v_0 is the within-trial variance of single simulations from the CSM. The judgment variance v is assumed to be constant across trials as well as the number of simulations k. The total variance is the sum of sampling variance and judgment variance. Thus, by fitting a linear regression from the variance of an individual CSM simulation (v_0) to the variance in participants' responses across all trials, we can estimate both $1/k$ and v. We show regression fits separately for **a)** Experiment 1 and **b)** Experiment 2; the dashed colored lines show the slope for different values of k. The results of this analysis show that the variances in participants' responses are most consistent with the assumption that they relied on $k = 1$ simulation to make their judgments (see also Battaglia et al., 2013; Vul et al., 2014).