Different Physical Intuitions Exist Between Tasks, Not Domains

Kevin A Smith[1], Peter W Battaglia[1,2], Edward Vul[3]

1: Department of Brain and Cognitive Science, Massachusetts Institute of Technology

2: DeepMind

3: Department of Psychology, University of California San Diego

Please send correspondence to:

Kevin A. Smith

Department of Brain and Cognitive Science

Massachusetts Institute of Technology

77 Massachusetts Avenue

Cambridge, MA 02139

Email: k2smith@mit.edu

Abstract

Does human behavior exploit deep and accurate knowledge about how the world works, or does it rely on shallow and often inaccurate heuristics? This fundamental question is rooted in a classic dichotomy in psychology: human intuitions about even simple scenarios can be poor, yet their behaviors can exceed the capabilities of even the most advanced machines. One domain where such a dichotomy has classically been demonstrated is intuitive physics. Here we demonstrate that this dichotomy is rooted in how physical knowledge is measured: extrapolation of ballistic motion is idiosyncratic and erroneous when people draw the trajectories, but consistent with accurate physical inferences under uncertainty when people use the same trajectories to catch a ball or release it to hit a target. Our results suggest that the contrast between rich and calibrated, versus poor and inaccurate patterns of physical reasoning exist as a result of using different systems of knowledge across tasks, rather than as a universal system of knowledge that is inconsistent across physical principles.

Keywords: domains of behavior, domains of knowledge, intuitive physics, rationality, heuristics

1 Introduction

Humans function remarkably well in varied, uncertain environments, but psychological research has documented many dramatic failures of human reasoning: we can walk over precarious terrain and stack dishes in elaborate arrangements in a drying rack, but we have trouble explaining how gravity works in basic situations (Hecht & Bertamini, 2000; McCloskey, Washburn, & Felch, 1983). Such discrepancies between robust, effective behavior and dramatic errors in simple problems have fueled key debates in behavioral economics (Camerer, 1987), communication (Piantadosi, Tily, & Gibson, 2011), reasoning (Tversky & Kahneman, 1983), and recently in the domain of intuitive physics (Marcus & Davis, 2013). Here we argue for a resolution to these tensions in intuitive physics: these differences in accuracy are not caused because we have a unified set of knowledge with large variance in accuracy across domains, but rather that we have different systems of knowledge – some more accurate than others – that we select depending on the task at hand.

People often make surprising errors in simple intuitive physics judgments such as drawing future trajectories of an object that has rolled off a cliff, has been dropped from a moving airplane, or released from a circular ramp (Caramazza, McCloskey, & Green, 1981; McCloskey, Caramazza, & Green, 1980; McCloskey & Kohl, 1983; Proffitt & Gilden, 1989; Ranney, 1994), but when people predict trajectories of billiard balls, estimate properties of colliding objects, determine how fluids will pour, or judge the stability of towers, their physical reasoning is often very accurate and consistent with the principles of Newtonian mechanics (Bates, Battaglia, Yildirim, & Tenenbaum, 2015; Battaglia, Hamrick, & Tenenbaum, 2013; Gerstenberg, Peterson, Goodman, Lagnado, &

Tenenbaum, 2017; Kubricht et al., 2016; Sanborn, Mansinghka, & Griffiths, 2013; Smith & Vul, 2013). Prior literature has attempted to explain this discrepancy by suggesting that some human knowledge of physical principles is accurate, while other knowledge is erroneous (e.g., people can estimate the stability of stacked objects, but have erroneous conceptions of ballistic motion; Marcus & Davis, 2013). This theory suggests that there are discrepancies across *domains* of physics: our system for intuitive physics includes accurate accounts of certain physical principles, but erroneous explanations of others.

However, there are also differences between these experiments in how this knowledge is measured: studies that demonstrate failures of physical knowledge tend to rely on asking participants to draw the future trajectory of objects or verbally explain how the world will unfold (e.g., diSessa, 1993; McCloskey et al., 1980; Shanon, 1976), while studies that show accurate knowledge tend to require people to make single judgments about a continuous physical property such as weight or future location (e.g., Battaglia et al., 2013; Sanborn et al., 2013; Smith & Vul, 2013). Differences in the literature might therefore be due to the types of tasks used rather than the types of knowledge required. Hegarty (2004) proposed that we employ different modes of physical reasoning depending on the format of the task, but to this date no one has directly tested for a difference in how people reason on tasks that require identical physical principles to solve but differ in how knowledge is queried. If intuitive physics is all derived from the same base of knowledge, we expect that people would demonstrate similar errors and biases across tasks that query knowledge in different ways; if people rely on different sources of knowledge, on the other hand, we would expect distinct patterns of behavior across different tasks.

In this paper, we test whether participants' behavior might differ between tasks that rely on the same physical principle using the classically-studied test-case of judging the ballistic trajectory of a ball released from a pendulum after the cord has been cut, and, if it does differ, how accurate physical reasoning is across tasks. In Experiment 1, participants each performed three distinct tasks: one *drawing* task, and two interactive tasks, *catching* and *releasing*. The *drawing* task replicated a classic failure of intuitive physics in which participants were shown static pictures of pendulums and asked to draw the path that a ball would take if the cord were cut at various points (Caramazza et al., 1981). In the *catching* and *releasing* tasks participants observed a pendulum in motion and were asked either to position a bucket to catch the ball once the pendulum cord were cut by a "knife" or release the ball from the pendulum so that it would be projected into a fixed bucket. All three of these tasks entailed solving the same physical problem – extrapolating the ballistic trajectory of a pendulum bob after the cord has been cut – so the systematic differences between human judgments in each task could arise only from the structure of the task itself, rather than differences in the underlying physical principles. We find that while people's drawings of such scenarios reveal behavior inconsistent with performance on the other two tasks, both catching and releasing predictions are consistent with the hypothesis that people form physical inferences using relatively accurate physical models perturbed by uncertainty (Battaglia et al., 2013; Sanborn et al., 2013; Smith & Vul, 2013).

In Experiment 2, we evaluated whether the differences between the tasks were due to the nature of the response format, or differences in the stimulus presentations. In the catching and releasing tasks, participants observed the pendulum in motion, but in the drawing task participants only observed a diagram of a pendulum on a sheet of paper. Because observing motion has been found to improve physical reasoning in certain tasks (Kaiser, Proffitt, & Anderson, 1985; Kaiser, Proffitt, Whelan, & Hecht, 1992), we investigated whether participants would demonstrate accurate physical knowledge on the drawing task only in the presence of pendulum motion. Although participants' judgments were different after observing motion, we find that these differences were driven by additional information about the velocity of the ball on the pendulum, but found no evidence that people used different, more accurate physical principles when they had access to richer stimulus information.

Together, these results suggest that differences in people's physical reasoning is not mainly driven by differences in domains of knowledge, but rather by the task they are solving. It is not the case that most poor performance is based on concepts within our physical knowledge that are categorically erroneous, but instead our capabilities differ based on the problem that we are using them to solve.

2 Experiment 1: Differences in behavior between tasks

2.1 Methods

2.1.1 Participants

Thirty-five UC San Diego undergraduates (with normal or corrected vision) participated in this experiment for course credit. All participants gave informed consent to participate in accordance with guidelines set by the UC San Diego Institutional Review Board. Participants were collected over a span of two weeks, and the number of participants was deemed appropriate before analysis based on pilot work (Smith, Battaglia, & Vul, 2013). Three participants were removed from analysis because their performance indicated that they were often responding randomly (see Supplemental Material, Figure S1 for details).

2.1.2 Procedure

Participants performed three blocked tasks that involved predicting the ballistic trajectory of a ball released from a pendulum: *catching*, *releasing*, and *drawing*. Participants always performed the drawing task after the other two tasks, but the order of the catching and releasing tasks was randomized across participants.

In the interactive tasks, participants viewed a computer monitor from a distance of approximately 60 cm, which initially depicted a ball swinging from a cord, consistent with pendulum motion. At some point in time the cord would be cut and the ball would be released, thus entering ballistic motion. A bucket was placed beneath the pendulum, and on each trial the participant's goal was to get the ball to drop into the bucket after being released. How they were allowed to interact with the scene differed between the catching and releasing tasks: participants could move the bucket but had no control over the point of release (catching) or could choose when to cut the cord to hit a fixed bucket (releasing). With the exception of one initial practice trial per task that familiarized participants with the scenario, the path of the falling ball was occluded in order to minimize learning. At the end of each trial, participants were given binary feedback that indicated whether or not the ball successfully landed in the bucket (we found no evidence of learning from this feedback; see Supplemental Materials section S2). A success earned participants a point, and each participant's score was totaled across all trials. This score was used solely as motivation to engage with the task and did not influence compensation or any of our analyses.

2.1.2.1 *Catching task*

Participants were instructed to adjust the bucket's horizontal position using the mouse so that the ball would land in the bucket after being released. The release time was pre-determined and varied across trials. Participants were notified of where the cord would be cut by an icon of a knife, which would darken when the cord was about to be cut. This knife let participants know where and in which direction the ball would be released from the pendulum so that they could begin forming their prediction before the cord was cut, thereby avoiding motor limitations from being unable to position the bucket quickly enough. The center of the bucket was recorded as the participant's judgment about where the ball would land (Figure 1, top; Movie S1).

2.1.2.2 Releasing task

The bucket was held fixed at a pre-determined position and participants were instructed to cut the pendulum cord by clicking the mouse at a time that would cause the ball to drop into the bucket. Cutting the cord was not allowed for an initial period of time randomly determined between 1.2-3.6 s to avoid biases from participants who would attempt to cut the cord as quickly as possible. The time at which the cord was cut was recorded for each trial (Figure 1, bottom; Movie S2).

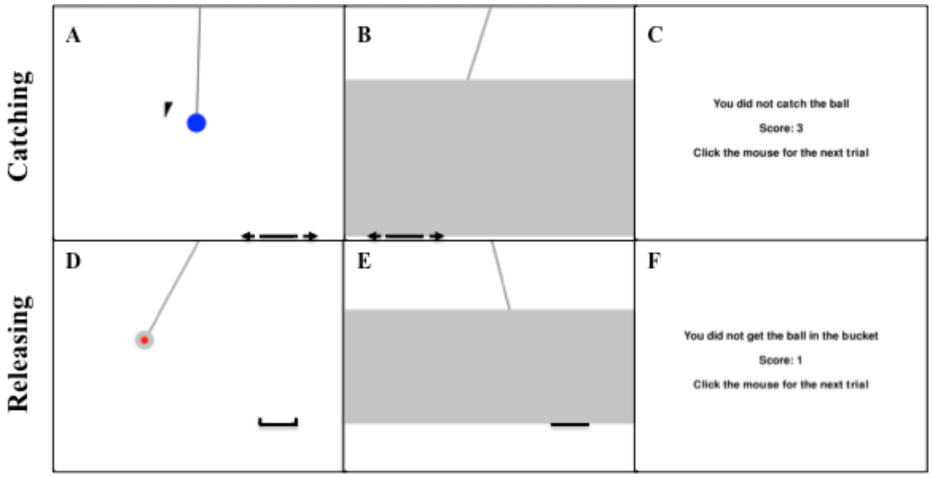

Catching
Releasing
A
B
C
You did not catch the ball
Score: 3
Click the mouse for the next trial
D
E
F
You did not get the ball in the bucket
Score: 1
Click the mouse for the next trial

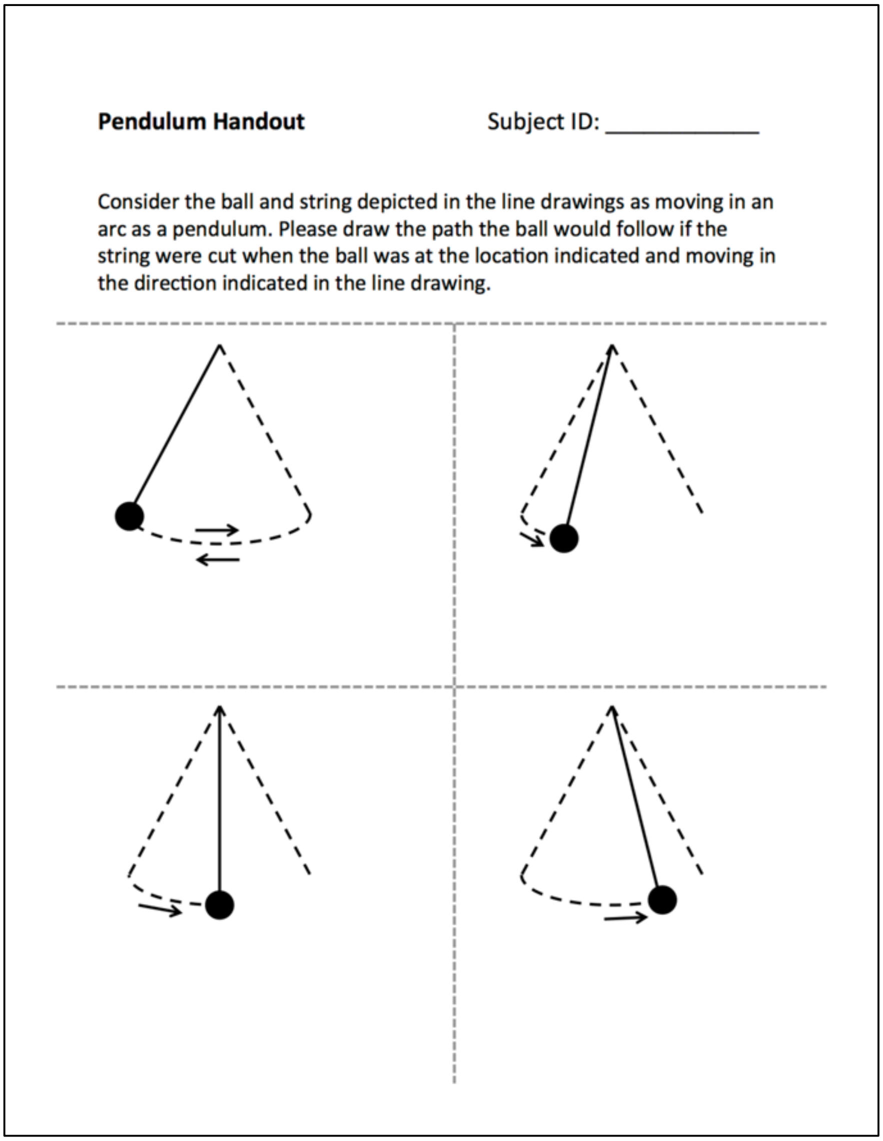
Figure 1: Diagram of trials in the catching (*top*) and releasing (*bottom*) trials. *Catching*: A. Participants observe a ball swinging on a pendulum and the 'knife' that will cut the cord. They move the bucket horizontally with the mouse. B. When the cord is cut, the trajectory of the ball is obscured. C. Binary feedback is provided after each trial. *Releasing:* D. Participants observe the ball on the pendulum. The coloring of the ball indicates a timer, such that once the red color is gone, participants can click the mouse to release the ball from the cord. E. The trajectory of the released ball is obscured. F. Binary feedback is provided.

For each of the these two tasks, participants repeated 48 different trials five times each in a randomized order. Trials were matched across tasks such that where the ball landed in a catching trial was the bucket position in the matched releasing trial. In the catching task, there were 16 distinct release times, crossed with three vertical distances between the nadir of the pendulum and position of the bucket – either 20, 35 or 50% of the total screen height.

Both tasks and all trials used the same pendulum. This pendulum had a length of half of the screen, and reached a maximum angle of 35° from vertical at its apex. The period (2.5s) and force of gravity were set to obey Newtonian mechanics as if the pendulum were positioned at a depth of 6m from the participants. This depth was selected to conform to participants' general expectations about the natural period of the pendulum as seen on the 2D computer screen used in the experiment, as determined by pre-experimental norming. To determine the motion of the pendulum, the cord was assumed to have negligible mass as compared to the ball, and so we could use simplified physical models to calculate the position of the pendulum at any point in time.

2.1.2.3 *Drawing task*

After the two computer-based tasks, participants were given a two-page packet to fill out. On the first page was the *drawing* task – a set of four diagrams depicting pendulums at different points in their swings. Participants were asked to draw the path that the ball would take if the cord were cut at the time depicted in the diagram (Figure 2). On the second page was a brief survey that asked about the participant's number of prior physics courses and strategies used in the experiment. These questions were reviewed to check whether participants were responding based on surface-level features or using other strategies that did not involve prediction – however, we did not find evidence of this.



Pendulum Handout Subject ID: ____________

Consider the ball and string depicted in the line drawings as moving in an arc as a pendulum. Please draw the path the ball would follow if the string were cut when the ball was at the location indicated and moving in the direction indicated in the line drawing.

Figure 2: Handout provided to participants for the drawing task. Instructions and stimuli were based on Caramazza et al. (1981).

To match drawing predictions to those from the catching and releasing tasks, we translated the drawing predictions to continuous measures by determining where participants would position the bucket in the catching task if their predictions were based on their drawings. To perform this translation, we first fit either linear or quadratic functions through the ball and the lines participants drew.[1] The first author and three research assistants at UCSD marked on each drawing at least five points along the drawn path, producing on average 54 points per drawing. We then fit two lines – linear and quadratic – through those points using least squares estimation. The quadratic fit was used for extrapolation unless either it had a positive quadratic term (implying the ball would move upwards), or the average distance between each point and the linear line was less than 1/8th cm more than the average distance from the quadratic line. In this way we allowed for curved drawings when appropriate but prevented inappropriate curvature that could bias results when the drawing itself was mostly linear. Additionally, this extrapolation smoothed out motor noise during production of the drawings. We then recorded where that extrapolated line crossed each of the three bucket heights, producing three "pseudo-catching" results per drawing – this yielded 12 results per participant.

In addition, similar to Caramazza et al. (1981), drawings were classified into one of eleven patterns that were either classifications from that experiment, or were observed in a pilot experiment (Smith et al., 2013). Three undergraduate research assistants from both UCSD and MIT who were naïve to the hypothesis performed this classification independently, and were told to match each participant's drawings to one of the given patterns as best as they were able, or rate the participant as *unclassifiable* if there was no matching pattern. A participant's drawing was considered matching a pattern if at least

[1] We could not simply determine where the drawn predictions crossed the line of each bucket height, since many drawings did not extend that far or ended at the left or right side of the drawing area. Therefore, we used a common extrapolation technique for all drawings.

two of the three raters agreed; if all raters disagreed, the participant's drawings were considered *unclassifiable*. There was high inter-rater reliability (*Fleiss' κ = 0.736*), and all three raters agreed on the classification for 23 of the 32 participants.

2.1.3 Models of physical reasoning

Even if people are using accurate physical principles to make predictions, uncertainty about the scene or motion of objects can cause biases in the predictions themselves (Battaglia et al., 2013; Sanborn et al., 2013; Smith & Vul, 2013). Therefore, to test whether participants were using accurate physical principles, we designed a model of physical prediction to determine how people would behave if they were basing their predictions on Newtonian mechanics. This 'calibrated' model assumes an idealized mental representation of the pendulum system, perturbed only by uncertainty about the depth location of the pendulum, and by accumulated noise in the trajectory of the ball throughout extrapolation (Smith & Vul, 2013).

We split this model into two parts: the predictive *forward model*, and the *task action*. The predictive forward model describes how a trajectory is extrapolated, and thus the physical understanding presumed under the model. For this task, this is instantiated as a prediction of where the ball will go when cut from the cord. This forward model is a set of rules shared across the catching and releasing tasks. We define the forward model as a function $R(t_{rel},y)$ which returns the predicted position where the ball would cross a line at height y if released at time t_{rel}.

The *task action* determines how those predictions are used to position the bucket or choose when to cut the cord, after incorporating noise/uncertainty from either accumulated prediction errors or motor control; these task actions differed across the two interactive tasks to account for the different ways of controlling the system (moving the bucket versus cutting the cord).

In addition to testing whether participants' predictions could be explained by calibrated physical reasoning, we also considered whether predictions could be explained by alternate, non-physical reasoning. To test these non-physical accounts, we compared variants with different non-physical forward models against one another (e.g., substituting the *R* function of the forward model); the task actions, however, stayed constant between models.

2.1.3.1 *Calibrated physics forward model*

The calibrated physics forward model assumes that people have an accurate knowledge of the laws of ballistic dynamics. To predict where the ball will land, the model uses Newtonian ballistic motion equations to extrapolate the path of the ball given its position and velocity at the moment of release, where [x_0, y_0] refers to the initial position, [v_{x0}, v_{y0}] refer to the initial velocity, t is the time since release, and g is the acceleration due to gravity:

$$x(t) = x_0 + v_{x0}t$$

$$y(t) = y_0 + v_{y0}t - \frac{1}{2}gt^2$$

Although we assumed people have good knowledge of the laws underlying the pendulum system, we also assumed participants were uncertain about the distance of the pendulum in depth from the observer – a necessary assumption since the 2D image of a pendulum on a computer screen is not interpreted as a physical pendulum literally at the depth of the computer screen. There is a lawful relationship between pendulum period, cord length, and the force of gravity that people are sensitive to (Pittenger, 1985), and participants directly observe the period of the pendulum and are assumed to have a sense of realistic Earth gravity (McIntyre, Zago, Berthoz, & Lacquaniti, 2001). But because the pendulum was presented on a computer screen with no depth cues, people must infer how far behind the screen they expect the pendulum to be positioned, and therefore the length of the pendulum cord. Because of the lawful relationship between pendulum length and gravity, changes in the depth of the pendulum have a direct correspondence to changes in gravity while holding the depth constant. Therefore for computational efficiency, our model assumed a constant cord length and estimated the effective strength of gravity (g) in px/s^2.

This means that calculating $R(t_{rel},y)$ involved determining the location $[x_0, y_0]$ and velocity $[v_{x0}, v_{y0}]$ of the ball at t_{rel}, then solving the differential equations above for the x-position given a specific y-position, assuming a positive t.

2.1.3.2 *Non-physical forward models*

Despite the body of literature studying the physical misconceptions that people hold, there is a dearth of formalized models about how people might understand ballistic motion. Most research has instead focused on conceptual descriptions of how gravity influences falling objects (Shanon, 1976) or how objects accelerate during their trajectory (Hecht & Bertamini, 2000). While Zago et al. (2004) suggest that people fail to account for gravitational acceleration in prediction, this only implies that non-physical models should predict that the ball will travel in a straight line. This account does not predict, however, the direction in which the ball should move when released. We therefore assumed that the extrapolations might be based on the same biased principles that have been found in previous drawing studies of ballistic motion (Caramazza et al., 1981). To do so, we formalized alternate forward models that could capture the same patterns participants made on the drawing task to test whether these patterns would be extended into the catching and releasing tasks. Each of these non-physical drawings can be captured by extrapolating the ball's path in a straight line at an angle away from the vertical (θ_r). For each model, this release angle was calculated as a function of the angle the pendulum cord made with the vertical at the moment of release (θ_c); however, the calculation of that angle varied by model (see Figure 3).[2]

In addition, Kozhevnikov and Hegarty (2001) suggest that people use impetus physics as their default implicit beliefs about physical events. Although no participants displayed beliefs of impetus physics in our drawing task (and only 11% of participants did so in Caramazza et al., 1981), we aimed to test whether an impetus model might capture participants' responses on the interactive tasks better than a calibrated physics model or other non-physical models. To formalize this model, we take the theory from Caramazza et al. (1981) and McCloskey (1983) that the ball will travel along the path of the pendulum (or beyond if near the apex) before losing all 'impetus' and falling down.

[2] Non-linear extrapolated trajectories, such as adding a quadratic term to the path, would make these non-physical models equivalent to a physical model of a parabolic ballistic trajectory; thus only linear extrapolation paths are guaranteed to differ from physical extrapolation.

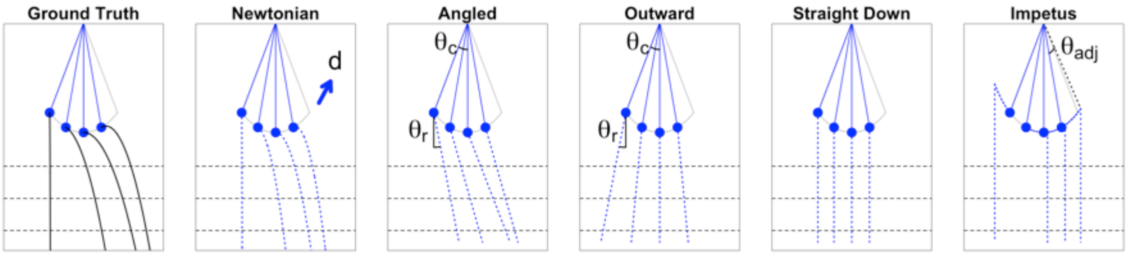


Figure 3: Diagrams of the forward model predictions for the path of the ball at four different cut points along the pendulum, using best fitting parameters. *Ground truth* is the path of the ball that was used to determine a successful catch in the experiment.

We therefore tested four non-physical models: *angled, outward, straight down*, and *impetus*. The *angled* forward model calculated the ball angle as a piecewise linear function of the angle that the pendulum formed with a line down its middle at release. There were two intercepts and two slopes for this function, so that the ball could travel differently depending on whether it was swinging downwards or upwards, and the angles were mirrored when the ball was travelling leftward for symmetry:

$$\theta_r = \begin{cases} i_1 + s_1\theta_c & \text{if } \theta_c > 0 \\ i_2 + s_2\theta_c & \text{otherwise} \end{cases}$$

The *outward* model assumed that the ball would continue along the path of the cord, but allowed for the angle to shift upon release. Thus the ball angle was calculated as the same as the release angle, with an adjustment *a* as a free parameter:

$$\theta_r = a * \theta_c$$

The *straight down* model simply assumed that the ball would drop upon release, which was equivalent to setting the release angle to 0:

$$\theta_r = 0$$

For each of the first three non-physical models, $R(t_{rel},y)$ was calculated by finding the ball's position [x_0, y_0] and the angle of the string θ_c at time t_{rel}, then using one of the equations above to calculate the release angle θ_r. Finally, the x-location where the ball would cross the given y-location was calculated by solving the system of equations:

$$x(t) = x_0 + t * \sin(\theta_r)$$

$$y(t) = y_0 + t * \cos(\theta_r)$$

Finally, the *impetus* model assumed that the ball would continue to travel a constant distance along the pendulum's arc in the direction of motion (possibly moving beyond the apex of the swing so that the ball could reach outside the horizontal confines of the pendulum arc), and then fall straight down. This formulation was chosen to match the diagrams representing impetus physics and participants' descriptions that "the ball will continue for a short time along its original arc, and then will fall directly to the ground" from Caramazza et al. (1981). Because the pendulum was identical in all cases, we instantiated this motion by adding a constant angular offset (θ_o) from the angle where the ball was released (θ_c) in the direction of the ball's motion, then assuming that the ball would fall straight down from that drop point (where L is the length of the pendulum):

$$\theta_d = \begin{cases} \theta_c + \theta_o & \text{if swinging rightward} \\ \theta_c - \theta_o & \text{otherwise} \end{cases}$$

$$x = L * \sin(\theta_d)$$

2.1.3.3 Task actions

The forward models provide a single deterministic prediction of where the ball will travel given that the cord is cut in a certain position, but people must use this information to interact with the task, specifically choosing where to place the bucket in the catching task or when to cut the cord in the releasing task.

In the catching task, participants observed when the ball was cut from the cord (r_{tr}) and the height of the bucket (y_{tr}), and were required to predict where it would land. This model captures human performance by using the forward model to determine where the ball should go given its release, and assumed that this would be the average location that participants would place the bucket. However, participants' responses were variable, and the model must capture this. Noise in tasks that require catching a hidden falling object includes both predictive and motor uncertainty (Faisal & Wolpert, 2009), both of which were modeled together as Gaussian noise around the predicted position. Since prediction error accumulates throughout the path, the model's uncertainty increases linearly with the vertical distance between the bucket and the release height of the ball (h_{tr}), where a_c and b_c are two free parameters to determine the linear fit:

$$\sigma_{tr} = a_c + b_c * h_{tr}$$

Thus the choice of where to place the bucket on the catching task (*S*) on a specific trial can be described as a normal distribution around the predicted landing spot according to the forward model:

$$S_{tr} \sim \mathcal{N}(R(t_{tr}, y_{tr}), \sigma_{tr})$$

In the releasing task, participants needed to solve the inverse problem from the catching task: given a specific landing position ($bucket_{tr}$, y_{tr}), where in the pendulum period should the ball be when the cord is cut? We assume that people always have a reasonable sense of where the ball will go if released at each point in time. This was approximated analytically by determining $R(t,y_{tr})$ for all t segmented in blocks of 10ms. From this information, we can form a function over possible release times that returns 1 if the ball will land in the bucket according to the model, and 0 otherwise:

$$L(t)_{tr} = \begin{cases} 1 & \text{if } R(t, y_{tr}) \in \text{bucket}_{tr} \\ 0 & \text{otherwise} \end{cases}$$

Assuming that motor errors are symmetric in time (Dawson, 1988), the optimal time to release the ball (T_{dec}) would be the middle of any contiguous period in which the ball would land in the bucket (e.g., L(t) = 1). If there were two contiguous periods of success,[3] we assumed that participants would be probabilistically more likely to choose the release point with the shorter vertical distance between the ball and the bucket (for a similar reason that we assume that uncertainty accumulates over vertical distance in the catching task). This choice was formalized as a logistic function on the difference between the ball heights at each point (*h*) with a single scaling parameter (s_r), but no intercept shift (since we assumed that at equal heights, participants should be ambivalent about which time to choose):

$$p(T_1) = \text{logistic}\big(s_r * (h_{T1} - h_{T2})\big)$$

Once the model chooses the time point that it aims to cut, its actual release time for a trial (T_{rel}) was selected as value from around that choice with Gaussian noise fit as a free parameter (σ_{terr}) to reflect the motor errors that people make:

[3] For instance, if the bucket were directly below the center of the pendulum, there are two periods when the ball could be released: when it is to the left of the bucket and traveling rightward, or when it is to the right of the bucket and traveling leftward.

$$\mathrm{T}_{rel} \sim \mathcal{N}(\mathrm{T}_{dec}, \sigma_{\mathrm{terr}})$$

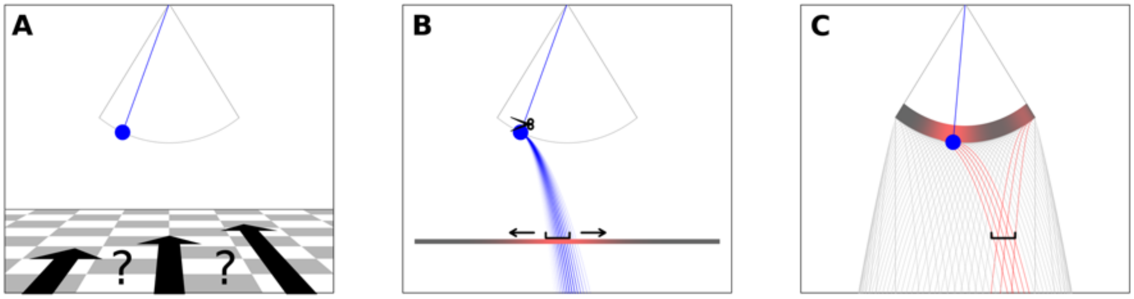


Figure 4: Illustration of the calibrated physics account of human judgments in the catching and releasing tasks. (*A)* Participants estimate the physical depth (and thus length) of the pendulum given its 2D projection and use this to guide both catching and releasing behavior. (*B)* In the catching task, participants see where the cord will be cut, generate noisy projections about where the trajectory of the ball will cross the plane of the bucket, and move the bucket into that region (red color mapping indicates higher probability of placing the bucket around that point). (*C)* In the releasing task, participants must choose when to cut the cord, so they project the ball's trajectory if released from different points spanning the pendulums' period, and choose a time to cut the cord that will make it probable that the ball lands in the fixed bucket given their motor timing error (red areas represent pendulum locations that are more likely to be selected as the release point).

2.2 Results

2.2.1 Consistency and accuracy of predictions

We tested how well each model captured human behavior by averaging over all predictions across all participants on a single trial – the average bucket placement for the catching task or average location where the ball hit the plane of the bucket for the releasing task – and compared that to an average of 500 noisy model predictions of the same measure. Participants' predictions in both of the interactive tasks were remarkably consistent with the calibrated physics model (catching: *r=0.988*, releasing: *r=0.998*, see Figure 5, left). Although participants' judgments were correlated with "ground truth" answers – responses under which the ball always landed in the center of the bucket (catching: *r=0.969,* cutting: *r=0.989*), judgments were systematically biased relative to ground truth (Figure 5, center). Moreover, these systematic biases were different between the catching and releasing tasks (error correlation of average bucket or landing position across matched trials: *r=0.35*), with participants typically positioning their buckets closer to the center of the screen in the catching task, and often overshooting the bucket in the releasing task (except near the edges, when they tended to undershoot the bucket). These unique task biases are expected under the calibrated physics model because each task reflects different sources of uncertainty subjected to the same non-linear transformation via Newtonian kinematics (model error correlation across matched trials: *r=0.19*); and indeed these systematic deviations of participants' judgments from the ground truth model matched the deviations of the calibrated physics model within each task (catching: *r=0.92*, releasing: *r=0.93*). By capturing these systematic biases, the calibrated physics model correlated better with participants' behavior than ground truth (catching: *z=2.26, p=0.02*, releasing: *z=4.28*, *p<.001*).

The positions where participants would place the bucket according to the imputed drawings were also well correlated with the calibrated physics model (*r=0.946*, Figure 5, bottom-left), but this appears to be due to the ability of the model to differentiate the four different drawing release points – this correlation is no different than comparing imputed drawings to ground truth (*r=0.946*, Figure 5, bottom-center). Furthermore, the errors between the imputed drawings and ground truth are anti-correlated with the errors from the calibrated physics model (*r=-0.41*), suggesting that human responses in the drawing task cannot be explained by assuming that people are using accurate physical principles.

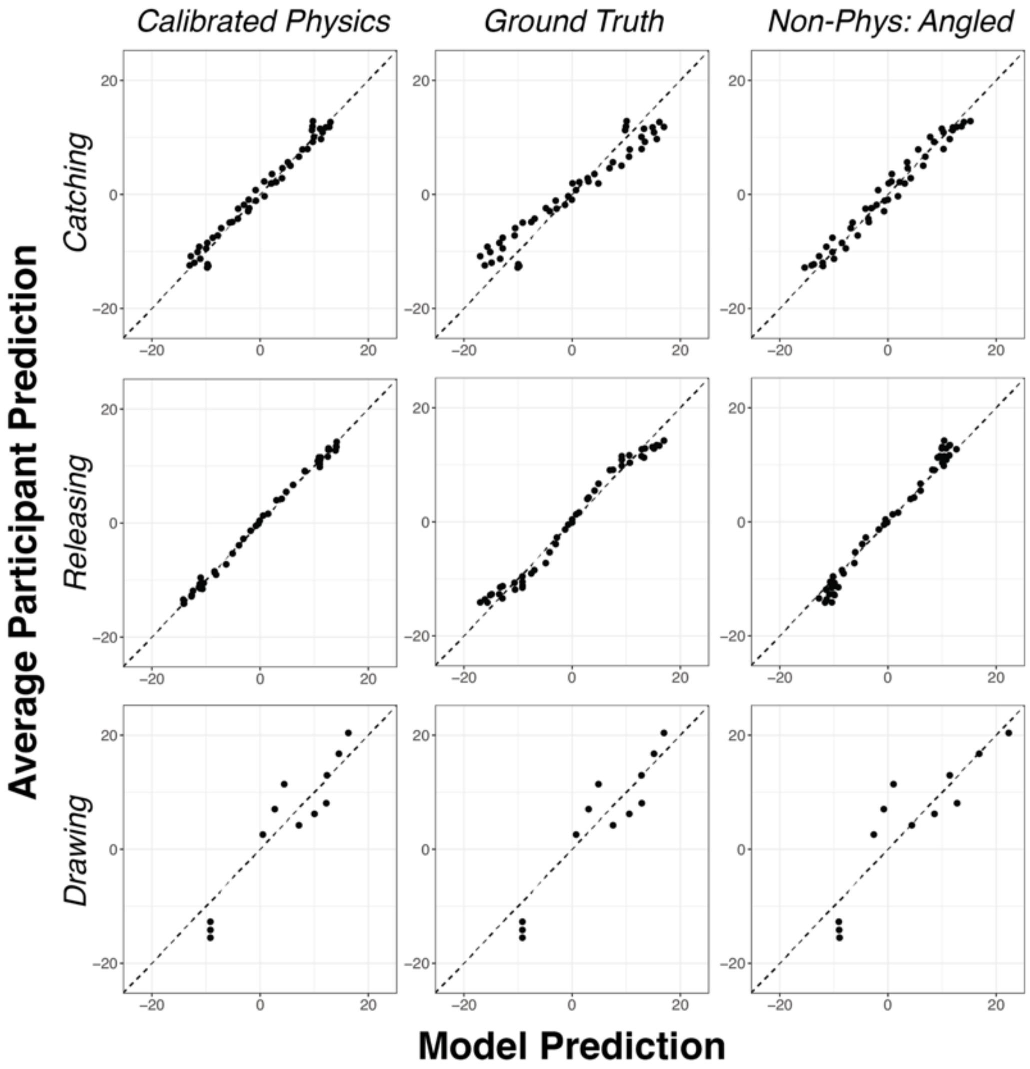


Figure 5: The bias and variance of participants' average performance on the interactive (catching and releasing) tasks is better captured by the noisy calibrated physics model than ground truth or the angled non-physical model. The drawings, on the other hand, could not be explained as well by calibrated physics. Each point represents one of 48 unique trials in either the catching or

releasing tasks, or one of the 12 imputed drawing trials. On the x-axis are model predictions (in cm from the center of the screen) of the position of the bucket (catching, drawing), the landing position of the ball if released at the predicted time (releasing), while the y-axis represents the average bucket (catching), landing position (releasing), or position where the imputed drawing path would cross the plane of the bucket (drawing) across all participants for that trial.

Behavior on the catching and releasing tasks also shows that non-physical forward models cannot explain human predictions. The calibrated physics model explained participants' catching and releasing responses better than any of the alternative forward models (angled: ΔBIC = 2,981; outward: ΔBIC = 14,435; straight-down: ΔBIC = 14,698; impetus: ΔBIC = 3,345; see Figure 6), suggesting that participants were not typically using an inaccurate heuristic to extrapolate the ball's motion.

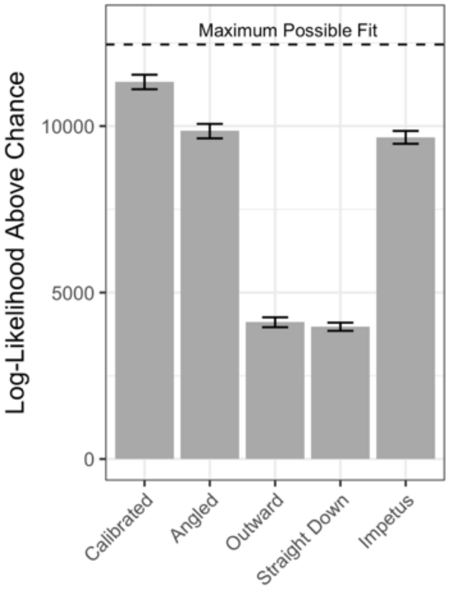


Figure 6: How well do the different models fit human catching and releasing behavior? The calibrated physics model explains participants' behavior better than any of the heuristic models. Log-likelihood above chance is the difference of the log-likelihoods of each of the models from

the log-likelihood of a random response model. Maximum possible fit is the log-likelihood of predicting behavior as well as possible from the behavior of other participants. Error bars are 95% confidence intervals, calculated from 500 bootstrapped samples each.

2.2.2 Individual physical knowledge

To test whether each participant was individually using accurate prediction, rather than such behavior arising only in the across-subject aggregate, we determined which of the calibrated physics and three heuristic models best described the behavior of each participant on the catching and releasing tasks. Of the 32 participants, 24 (75%) were best fit by the calibrated physics model, and 8 (25%) by non-physical models based on BIC.

Because of the sparsity of data from the drawing task, we could not reliably fit individual models to participants' drawing data. Instead, we classified drawings in a similar way to Caramazza et al. (1981), using raters to match drawings to a series of potential patterns (see Section 2.1.2.3, Figure 7). Unlike the catching and releasing tasks, only 7 (22%) of participants were observed to draw paths roughly consistent with calibrated physics.

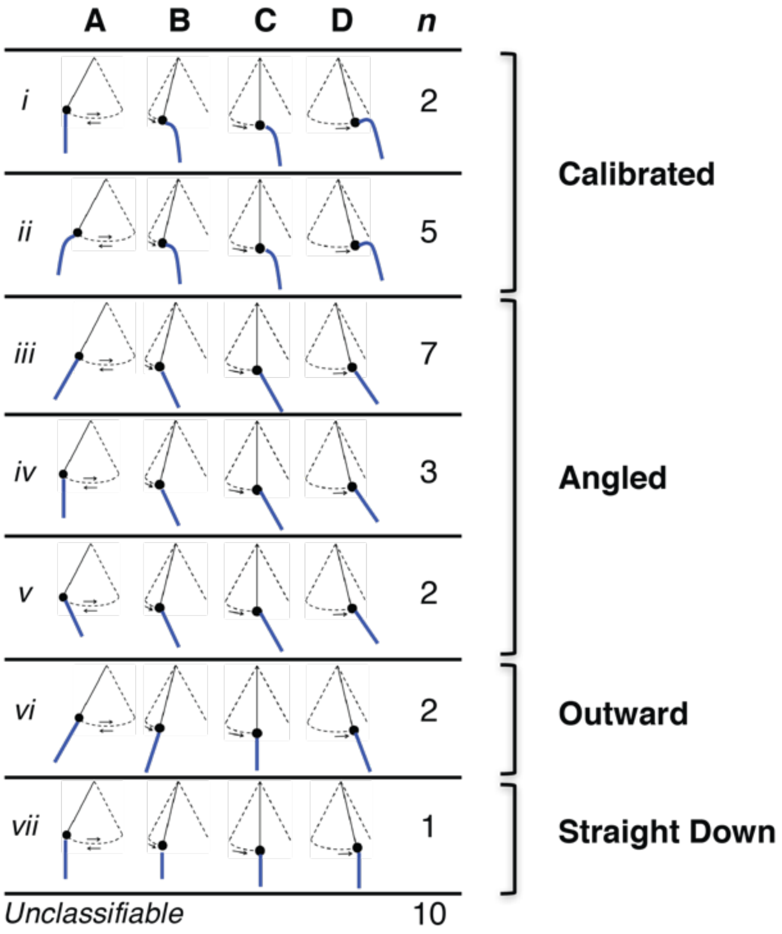


Figure 7: Classification of responses on the drawing task, grouped by type of forward model that could generate them. Few (6%) participants drew perfectly accurate paths for all four diagrams (classification *i*), suggesting that most participants are not using calibrated physical principles for performing this task.

Participants' drawings were also inconsistent from person to person: no more than 22% of participants were classifiable into a single category of response patterns. This drawing variability mirrors behavioral variability in similar physical tasks (Caramazza et al., 1981; Kaiser, Proffitt, & McCloskey, 1985; Proffitt, Kaiser, & Whelan, 1990). This idiosyncrasy is highlighted by how well we can predict participants' response errors from the average errors of all other participants: if participants are all biased in a similar fashion, it suggests that they are relying on similar cognitive processes. Each participant's error was highly correlated with the errors of other participants in the catching (*mean r=0.76, 10-90% quantiles =[0.47,0.91]*) and releasing tasks (*mean r=0.53, 10-90% quantiles =[0.33, 0.64]*, see Figure S2), which is consistent with most participants relying on a similar mental model to that of other participants. In contrast, the correlation between each participant's imputed drawing errors and that of other participants was lower (*mean r=0.29, 10-90% quantiles =[-0.66, 0.89]*; see Figure S2). This suggests that the drawing task either relies on a much noisier read-out from the same process, or that people use more idiosyncratic processes on the drawing task.

Moreover, there were inconsistencies between individual participants' behavior on the catching and releasing tasks, and the paths produced in the drawing task. We first tested whether participants' imputed drawings could be predicted from their responses on the matched trials of the catching task. For each participant, we took the average responses across the 12 catching trial types that were matched to the imputed drawings, and asked whether those predictions reliably correlated with that participant's imputed drawing path; however, we found no reliable evidence of this correlation across participants (*mean r=0.17, 10-90% quantiles = [-0.43, 0.68]*). Because this analysis relied on a sparser set of trials, we also checked whether we could reliably predict a relationship we expect to exist: predicting a single catching response from other responses in the same situation. Since each participant made judgments on the same catching trial type five times each, for the same 12 trial types we tested whether the average of four of those predictions reliably correlated with the held out prediction on the same trial, making this as analogous to the comparison with imputed drawing predictions as possible. We randomly segmented predictions and calculated the correlations 100 times for each participant, and did find a reliable correlation across participants (*mean r=0.52, 10-90% quantiles = [0.22, 0.84]*, all *rs > 0*), and these correlations were statistically higher than the imputed drawing correlation (paired Wilcoxon test: *V=54*, *p=0.0025*).

Furthermore, if we look at how the catching and releasing models compare to the classification of the drawings, none of the eight participants fit best by non-physical forward models on the catching and releasing tasks had drawn extrapolated trajectories consistent with the heuristic model that best captured their interactive task behavior (see Table 1).[4] These results suggest that the population does not contain subsets who have universally incorrect knowledge of physics across cognitive domains. Instead, when interacting with physical scenes, people share a common system of physical knowledge, calibrated with the world, while their drawings of trajectories in those same scenes may be guided by idiosyncratic and often non-Newtonian heuristics.

[4] Even if the 'impetus' model is not included (because no participants drew diagrams consistent with impetus physics), there are still no participants who shared a non-physical classification between the interactive and drawing tasks. All except one of the participants who were best fit by the impetus model would be best fit by the calibrated model if the impetus model was not included, and that participant drew a calibrated path but was best fit by the angled model.

		Model Fit				
		Calibrated	**Angled**	**Outward**	**S. Down**	**Impetus**
Drawing Task	**Calibrated** (*i, ii)*	**5**	**0**	**0**	**0**	**2**
	Angled (*iii, iv, v*)	**12**	**0**	**0**	**0**	**0**
	Outward (*vi*)	**1**	**1**	**0**	**0**	**0**
	Straight Down (*vii*)	**1**	**0**	**0**	**0**	**0**
	Unclass.	**5**	**1**	**1**	**0**	**3**

Table 1: Best fitting model to joint catching/releasing predictions vs. classification on drawing task for each individual participant. Roman numerals refer to the drawing type classification from Figure 7. No participant was best fit by a non-physical model that could capture his or her drawing classification.

3 Experiment 2: The impact of stimulus richness on physical knowledge

We found in Experiment 1 that interactive tasks tapped into relatively accurate models of physical reasoning, while participants relied on idiosyncratic and potentially erroneous physical reasoning to solve the drawing task. However, the tasks in Experiment 1 differed not just in the way that we queried participants' knowledge, but also in the information available to participants to perform the tasks: in the catching and releasing tasks, participants observed the pendulum in motion, while in the drawing task participants were given a sheet of paper displaying a static pendulum. Prior work has suggested that viewing moving stimuli can produce more accurate physical judgments (Kaiser, Proffitt, & Anderson, 1985; Kaiser et al., 1992). However, these experiments contrast full dynamic information – showing the motion of the pendulum system both before and after release – with static pendulums followed by choices between line trajectories, and with "kinematic" trajectories that follow the path of the line drawing at a constant speed (e.g., not accelerating naturally due to gravity). Crucially, they find that only in the full dynamic condition do participants choose the correct trajectory more often.

Thus there are two possibilities for why performance in these experiments with moving stimuli is better: either it is because the nature of the response itself differs (e.g., it is easier to compare a mental simulation to a dynamic movie than to a static diagram, so simulation might be more preferred for dynamic stimuli), or because viewing the motion prior to release makes simulation using a calibrated model more likely (e.g., it is easier to create a mental model to simulate when additional dynamic information is available), or both. To tease these possibilities apart, we test how people perform the drawing task with a moving pendulum to determine whether participants with motion information would rely on more accurate physical principles than those that must rely on static stimuli.

Although showing moving pendulums does change the predictions that people make on the drawing task, these changes are not due to people using more accurate physical principles but rather from making different inferences about the velocity of the ball at the moment the pendulum string is cut.

3.1 Methods

3.1.1 Participants

Sixty-seven UC San Diego undergraduates (with normal or corrected vision) participated in this experiment as part of a set of experiments for course credit. All participants gave informed consent to participate in accordance with guidelines set by the UC San Diego Institutional Review Board. We collected data until we had approximately twice the number of participants from the original task. Participants were randomly assigned to the *Motion* or *Static* conditions, resulting in 33 participants in the Motion condition and 34 participants in the Static condition.

3.1.2 Procedure

Participants were instructed that they would need to judge the path of a ball that is cut from a pendulum, and that they would indicate the ball's predicted path by clicking and dragging the mouse. Participants in the *Motion* condition observed the pendulum make one full swing then swing to the point where the string would be cut, while participants in the *Static* condition observed only the final position of the pendulum as the string is cut; therefore participants in both conditions observed identical images immediately before being asked to respond. The pendulum used here was identical to the pendulum used in the *Catching* and *Releasing* tasks, with the same arc and, for the *Motion* condition, the same period.

Participants indicated their predictions by clicking and dragging along the path they believed the ball would travel. To ensure that we captured paths of appropriate length, these paths were required to (a) start from within the image of the ball and (b) terminate within 10% of the edge of the lower half of the screen; if the path did not meet these criteria, participants were notified and asked to draw the path again. Finally, participants would be asked to either confirm their path, or click a ‘Try Again’ button to re-draw it (see Figure 8: top).

All participants drew their predictions for the same four release points measured in the *Drawing* part of Experiment 1; the order of presentation was randomized across participants. For each drawing, we recorded each point along which participants dragged the mouse, measured every 20ms, from which we could reproduce the drawn path.

3.1.3 Rating

As with the *Drawing* task of Experiment 1, we asked three undergraduate raters from UCSD to classify each participant’s drawings. Because we hoped to test how judgments varied in detail, we asked the raters to judge the predictions individually by stimulus, rather than the pattern of predictions across all four stimuli. Raters classified each drawing into one of six types (see Figure 8: bottom), or judged an individual drawing to be *unclassifiable*. Raters were blind to which participant created each stimulus and to whether they were in the *Motion* or *Static* condition.

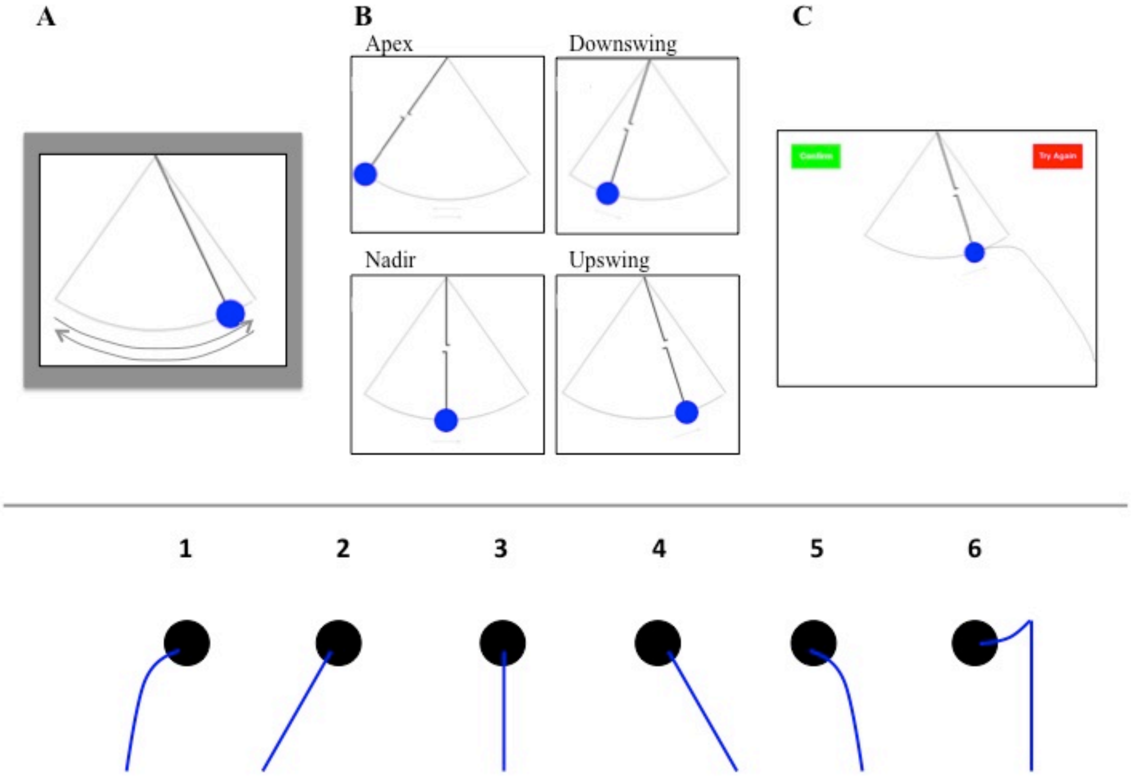


Figure 8: *Top:* Diagram of a trial. A: participants in the *Motion* condition only observed the pendulum swing through one full period, then swing to the final position. B: participants in both conditions would observe a static image of the pendulum string cut at one of four positions. C: participants click and drag the mouse to indicate their predictions for the ball's motion. *Bottom*: The six potential paths raters could classify each drawing as (not including *unclassified*). All of the patterns from Experiment 1 or Caramazza et al. (1981) could be recreated from a combination of these path types.

Inter-rater reliability was lower than the reliability from Experiment 1 (*Fleiss' κ = 0.596*), but this effect was driven by one rater who had a higher threshold for classifying drawings (rating 42% of drawings as *unclassifiable*). Reliability where this rater classified drawings was very high (*Fleiss' κ = 0.826*), and on the stimuli she determined to be unclassifiable the other two raters agreed on a classification 79% of the time. Similar to Experiment 1, we classified each drawing as the majority classification of the raters, but if all three raters disagreed, we noted the drawing as *unclassifiable* (this was only true of 5% of the drawings).

3.2 Results

3.2.1 Differences in predictions by condition

We first tested whether there was evidence of differences in participants' predictions due to motion evidence for each pendulum cut point. If motion information does not affect physical reasoning, then we should expect no difference between participants' predictions in the *Motion* and *Static* conditions. On the other hand, if motion information causes people to use accurate models of physics, then participants in the *Motion* condition should make different predictions from those in the *Static* condition, and should draw more curved paths to indicate the appropriate influence of gravity on the ball's ballistic trajectory.

We did find evidence that participants' drawings differed between the two conditions in for both the Apex ($\chi^2=13.4, p_{sim}=0.014$) and the Nadir ($\chi^2=10.7, p_{sim}=0.035$) pendulums, but not in the Downswing ($\chi^2=2.4, p_{sim}=0.71$) or Upswing ($\chi^2=8.3, p_{sim}=0.14$) stimuli. The differences in Apex predictions appear to be driven by participants with motion information believing that the ball retains leftward velocity, while participants without motion information tend to believe the ball will drop (the correct answer) or travel to the right. The difference in Nadir predictions are driven by participants without motion information indicating that the ball will drop straight down, while participants with motion information realize that the ball retains horizontal velocity (see Table 2).

Although there is evidence that motion does influence peoples' predictions, there is no evidence that it causes them to use more accurate physical principles for those predictions. For the Downswing, Nadir, and Upswing stimuli, there was no evidence that participants in either condition drew the correct ball path at different rates (path 5 from Figure 7; all $\chi^2 < 0.5$, all $p_{sim} > 0.5$). There was a difference in accuracy with the Apex condition, but it was participants in the *Static* condition who were more likely to be correct (24% vs. 6%; $\chi^2 = 4.4$, $p_{sim} = 0.043$). Thus pendulum motion provides different information about the ball's velocity, but this information can be misleading (e.g., indicating the ball retains velocity at the apex) and does not cause people to produce more correct parabolic paths.

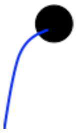
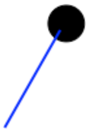
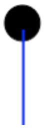
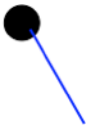
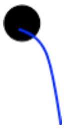
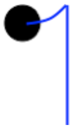
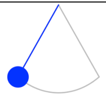
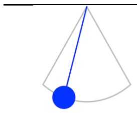
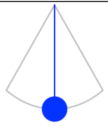
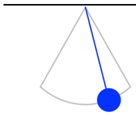
		1 	**2** 	**3** 	**4** 	**5** 	**6** 	**Unclassified**
 **Apex**	Static	5	1	8	10	1	0	8
	Motion	10	9	2	5	1	0	7
 **Downswing**	Static	0	2	2	18	7	0	6
	Motion	0	2	0	19	7	0	4
 **Nadir**	Static	1	1	13	7	7	0	8
	Motion	0	0	5	16	5	0	4
 **Upswing**	Static	0	7	1	8	8	3	6
	Motion	0	0	2	10	10	5	6

Table 2: Classification of participants' drawings, split by pendulum cut point and experimental condition. The veridical response was 5 in all cases except the Apex, where the veridical response was 3. Patterns of responses between the *Static* and *Motion* conditions differ in the Apex and Nadir scenarios based on differences in how participants interpret the ball's velocity, but there is no evidence that the physical principles used differ between conditions.

3.2.2 Consistency of Static and Motion predictions

The drawing classifications demonstrated that gross visual features of drawings differed between conditions only for the Apex and Nadir stimuli, but we further tested whether this was due to differences or similarities in overall predictions between participants in the *Motion* and *Static* conditions, or whether this was simply a difference in the features of the trajectories they drew.

We extrapolated the drawings in the same way as Experiment 1 as a separate test of how consistent the *Motion* and *Static* predictions were (see Section 2.1.2.3).[5] If people were making different predictions based on the motion information, then we would expect that any individual's imputed drawings should be better explained by participants' prediction from the same condition than from the opposite condition; conversely, if behavior can be equally well described by the predictions of both condition groups, it suggests that there are no differences in the overall predictions between the two groups. We therefore test how well the errors (as compared to ground truth) that each participant makes correlate with the average errors from participants in the same or the other condition. Because we expect large variability in individual's drawings and have sparse data, any single correlation would not be informative, but averages across participants from each of the two conditions can suggest that, on average, people make similar or different predictions from others given the same motion information.

Similar to Experiment 1, participants drawing errors were not well correlated with the average errors from all other participants and were extremely variable (*mean r=0.29, 10-90% quantiles =[-0.62, 0.89]*). However, this did vary as a function of condition: extrapolated drawing errors from the *Static* condition were more correlated with other *Static* errors than *Motion* errors (Static: *mean r=0.50, 10-90% quantiles =[-0.13, 0.89]*; Motion: *mean r=0.02, 10-90% quantiles =[-0.39, 0.47]*), while the *Motion* extrapolation

[5] Because we captured points along the drawn line as part of the task we did not have third parties mark each drawing, but the technique for extrapolating lines from the drawn points was identical.

errors were somewhat more similar to other *Motion* errors than *Static* (Static: *mean r=0.07, 10-90% quantiles =[-0.78, 0.94]*; Motion: *mean r=0.16, 10-90% quantiles =[-0.40, 0.65]*).

However, this effect was driven almost exclusively by differences in prediction for the Nadir cut point; excluding this stimulus, participants' errors in the *Static* condition were equally well correlated with the average errors in both conditions (Static: *mean r=0.45, 10-90% quantiles =[-0.49, 0.91]*; Motion: *mean r=0.42, 10-90% quantiles =[-0.12, 0.87]*), as were participants in the *Motion* condition (Static: *mean r=0.17, 10-90% quantiles =[-0.80, 0.95]*; Motion: *mean r=0.22, 10-90% quantiles =[-0.67, 0.81]*).

This provides further evidence that seeing the pendulum in motion provides additional information about the velocity of the ball at the moment that the string is cut: for the apex and nadir stimuli predictions do differ between groups as a result of this motion information, but we do not have evidence that people are using different processes to produce their drawings for the downswing and upswing stimuli.

4 Discussion

Across two experiments we asked people to make physical judgments in several different tasks, all of which depended on identical underlying physical principles. In Experiment 1, participants used relatively accurate principles to predict the ballistic trajectory of a ball cut from a pendulum, but were idiosyncratic and inaccurate when drawing that trajectory. In Experiment 2, participants continued to use erroneous physical principles for drawing trajectories, even with richer, less abstract stimulus information.

4.1 Differences in intuitive physics systems by task

Across these experiments we demonstrate that it is not simply the case that people have accurate internal models of some physical principles and inaccurate models of others; instead, the knowledge we bring to bear depends on the requirements of the task at hand. This raises the crucial question of what causes this difference between tasks.

4.1.1 Cognitive systems for physical reasoning

Addressing why we see different behavior by tasks requires us to hypothesize what cognitive systems underlie behavior in the current experiments so that we can assess how features of each task might drive the use of different systems. Behavior on both the catching and releasing tasks was best explained by a model of physics that approximates Newtonian mechanics perturbed by uncertainty – a hallmark of the "intuitive physics engine" proposed by Battaglia et al. (2013). This theory suggests that we build mental models of the world that we can simulate forwards to predict how the world will unfold. On the other hand, previous studies that have found erroneous conceptions of ballistic motion have asked participants to produce (McCloskey, 1983) or assess (Hecht & Bertamini, 2000) verbal descriptions of events. These judgments therefore may be formed from discrete, verbalizable atoms of knowledge (diSessa, 1993) that can be combined into logical or rule-based explanations and decisions (e.g., "the ball will fall downwards and to the right"). In support of this theory, Caramazza et al. (1981) found that biases are attenuated by formal instruction which suggests that this cognitive process can be easily changed by verbal information. To summarize, the "intuitive physics engine" requires rich information about the world but can provide more precise, quantitative information, whereas rule-based systems can handle less specified world models, but cannot provide as precise outcomes (Davis, Marcus, & Frazier-Logue, 2017).

4.1.2 Prior accounts of simulation versus logic

Previous work has suggested cases where simulation-based physical reasoning is prioritized over logical or analytic strategies, but these explanations cannot capture the differences observed in the current experiments. For instance, Schwartz and Black (1996) find that when people are given realistic pictures of a physical system, they are more likely to use mental simulation to answer questions about that system; conversely, with abstract diagrams they are more likely to use analytic processes. However, the pendulums observed in the drawing task of Experiment 2 were identical to the pendulums of the catching and releasing tasks of Experiment 1, and so in this case it was only a difference of response modality and not stimulus realism that drove the difference in behavior.

Kozhevnikov and Hegarty (2001) suggest that when people must make immediate responses they rely on default, implicit beliefs, but when they have a chance to reflect, they can override these intuitions with explicit, verbalizable knowledge. However, our findings are not captured by this framework either. First, Kozhevnikov and Hegarty (2001) find that implicit knowledge is erroneous and explicit knowledge can correct these misconceptions, whereas we find more accurate physical principles from the system that would map onto the “intuitive” beliefs. Second, both the catching and releasing trials were not immediate – they took a few seconds to resolve each – and yet we found no evidence that any of our participants were systematically using the same information they used on the drawing task.

4.1.3 Metareasoning over cognitive systems

Instead, we hypothesize that the cognitive system people use for physical reasoning is chosen based on both (a) the ease of recruiting analog, simulation-based versus logical, rule-based mental models based on the way the stimulus is presented, and (b) the information that must be produced to solve the task at hand. Choosing a cognitive system can be construed as metareasoning: deciding how to deploy cognitive resources in an efficient manner, considering both the expected utility and expected costs rather than using the most accurate strategy regardless of resource usage (Russell & Wefald, 1991). This framework has been shown to explain general human strategy selection (Lieder & Griffiths, 2017; Payne, Bettman, & Johnson, 1988), and has been used to describe how people allocate cognitive resources within a strategy for physical prediction (Hamrick, Smith, Griffiths, & Vul, 2015).

While past work (e.g., Schwartz & Black, 1996) has focused on how the stimulus presentation affects strategy choice, we focus instead on the differences that arise by task. Here, both the catching and releasing tasks require a precise, continuous response to get the ball in the bucket, since a small difference in the bucket position or release time could be the difference between a correct and incorrect answer. On the other hand, the pragmatic implication of asking people to "draw the path of the ball" might require less precision. Participants were not given a metric of success against which their drawings would be measured, and therefore may assume that the experimenter is assessing their predictions in a less strict fashion: it might be unreasonable to believe that the experimenter will match up the drawing against the exact trajectory, but reasonable to think that they only need to communicate the general direction that the ball will travel after release. This would, for instance, explain why participants often did not draw trajectories to the edge of the diagram in the first experiment, necessitating the requirement that drawings reach the edge of the screen in Experiment 2. In this case, it might be cognitively costly to read out multiple points from a simulated trajectory, whereas it is easy to use a simple logical analysis of the scene to produce a drawing (e.g., "the ball has rightward velocity and gravity will make it fall, so it will travel down and to the right"). Thus using simulation versus logical analysis might provide a greater expected benefit for the catching and releasing tasks, but the opposite could be true for the drawing task. An important area for future research would be to directly test this theory by, for instance, changing the level of precision implied in the drawing task.

Framing the choice of intuitive physics systems as a metareasoning problem can also provide an alternate account of why people are better at judging accurate ballistic trajectories in the presence of full dynamic information versus static diagrams – even though in theory both should require knowledge of the full trajectory of the object. Prior work has explained this finding by positing a perceptual system that can differentiate natural from non-natural motion, but suggests that people cannot easily use this system for other judgments about dynamic stimuli (Kaiser, Proffitt, & Anderson, 1985). Yet if people are simply asked to imagine a pendulum swinging before making explicit judgments about its motion, their judgments are more accurate (Frick, Huber, Reips, & Krist, 2005), suggesting that this information can in fact be made accessible for more explicit judgments. We propose that there is no special system for judging perceptual naturalness, but instead that naturalness judgments happen to be well-suited for the output of an analog simulation system in a way that judging static trajectories is not:

naturalness judgments require tracking and matching the precise position of an object over time, and therefore requires more precision than producing a trajectory in the same scene or comparing static trajectory diagrams that do not have the same motion over time. Future research is needed to disambiguate these two hypotheses, but investigating task differences could explain *why* we find differences across the two types of tasks.

4.1.4 An alternate account: one system, differential noise

An alternate hypothesis to tasks relying on multiple systems, however, is that the drawing task relies on a noisier readout from the same intuitive physics engine as the catching and releasing tasks, perhaps because extracting multiple points leads to an autocorrelation bias, or because of the pragmatic implication that less precision is needed. In this case, a few noisy queries from a parabolic path might be described in a roughly linear way, which would show up both in drawing classification and as errors in extrapolation. This theory would still suggest that responses should differ by task, but differences should arise from noise in responses rather than a difference in cognitive processes. However, this theory would also conflict with prior interpretations of intuitive physics results: people's verbal descriptions of object motions often contain the same errors as their drawn trajectories (McCloskey, 1983; McCloskey et al., 1980), which is taken as evidence of using the same principles to produce both drawings and descriptions. If the drawings are just noise-corrupted read-outs from an accurate physical model, then these explicit descriptions could only be post-hoc reasoning that fits the previously produced drawings; yet people still produce erroneous descriptions even without drawing a trajectory first (Hecht & Bertamini, 2000; Shanon, 1976). However, this theory would require that naïve explanations of physics are all based on noisy readouts from our intuitive physics engine, rather than logically constructed from memory or more atomic pieces of information as previously proposed (diSessa, 1993). Thus in light of prior research into the construction of physics explanations, it is more likely that people are using two cognitive systems for intuitive physics rather than one system with differential noise.

4.2 Systems of reasoning

These findings mirror a broader pattern of results in the psychological literature: people's behavior differs between tasks that require interaction with the environment, and those that require verbal responses (Chen, Ross, & Murphy, 2014; Glaser, Trommershäuser, Mamassian, & Maloney, 2012; Wu, Delgado, & Maloney, 2009). Some behavior, especially in lower level perceptual and motor domains, is near optimal given the information and processing constraints associated with a particular task (Griffiths & Tenenbaum, 2006; Stocker & Simoncelli, 2006; Trommershäuser, Landy, & Maloney, 2006; Wolpert, Ghahramani, & Jordan, 1995) while other behavior, especially in higher level cognition, is subject to gross biases and errors (McCloskey et al., 1980; Tversky & Kahneman, 1983). This dichotomy mirrors a well-known theory in the decision making literature: our intuitive decisions are often thought to be based on a different system of reasoning than deliberative choices (System 1 vs. System 2; Kahneman, 2011).

Often these dichotomies are characterized as 'automatic systems' (System 1) and 'deliberative systems' (System 2), and this split may be appropriate for physical reasoning as well. When we are throwing a ball to a friend we are unaware of the complex calculations that must be done to determine the exact force and angle that we will throw the ball with, yet when we solve simple high school physics problems many of us are well aware of the effort that it can take. Thus our ‘intuitive physics engine’ (Battaglia et al., 2013) may be the analog to the ‘automatic’ system – one that works in a cognitively impenetrable way and perhaps is brought online without conscious effort in the presence of suitable physical motion (Fischer, Mikhael, Tenenbaum, & Kanwisher, 2016). But in addition, we can also use other verbalizable knowledge to ‘deliberatively’ structure explanations and descriptions of events (diSessa, 1993).

4.3 Intuitive physics for action

While we propose that participants use two different systems in these experiments, this is not meant to be exhaustive. For instance, Zago and Lacquaniti (2005) further differentiate between *perceptual* knowledge and *motor* knowledge, suggesting that we rely on different cognitive systems for, e.g., determining when a falling ball will cross a line versus catching that ball at the same point (Zago et al., 2004). This is similar to the visuomotor control literature that shows that the motor system is not biased by the same illusions that affect perception (e.g., Aglioti, DeSouza, & Goodale, 1995), suggesting a distinction between “vision for perception” and “vision for action” (Glover, 2004; Goodale & Milner, 1992).

Similarly, there may be a distinction between “physics for perception” and “physics for action”, where motor control for, e.g., grasping moving objects relies on a separate cognitive system (Zago et al., 2004). The experiments in this paper involved interacting with the objects via a computer mouse rather than directly manipulating them and so would rely on a perceptual system according to this distinction; however, it will be important in future work to map out whether there is a shared or different intuitive physics engines for perception and action.

4.4 The accuracy of the intuitive physics engine

We found that participants' behavior when interacting with the scene in the catching and releasing tasks was consistent with accurate physical principles. This contrasts with the theory of Kozhevnikov and Hegarty (2001) who suggest that "implicit" knowledge of physics is erroneous but could be corrected by explicit knowledge, and the theory of Zago and Lacquaniti (2005), who propose that that perceptual physics knowledge is biased but our motor predictions use calibrated physics. The theories suggesting biased "implicit" physics have been based on evidence from the "representational momentum" literature that memory for the location of objects is shifted based on the dynamics of impetus physics rather than Newtonian mechanics (e.g., if two objects are moving downward on the screen then disappear, people will remember the larger object as having "fallen" further than the smaller object; Hubbard, 1997). However, it is unclear whether these representational momentum phenomena arise from object dynamics or from perceptual biases (Kerzel, 2002), and therefore these findings may not be applicable to our implicit physical reasoning. Zago et al. (2004) additionally suggest that our perceptual predictions do not account for gravity because people's timing to press a button in response to a falling ball passing a marker appeared to be unaffected by gravitational acceleration. Here we did not measure *when* participants believed that the ball might hit the bucket, which might suggest that the timing estimates in physical prediction might be distorted compared to reality. However, the results here suggest that we do take the curvature of gravity into account to extrapolate the path we believe the ball will take.

Note that this claim of accuracy is not a claim that we explicitly solve Newtonian equations whenever we predict the motion of an object, or even that our predictions are perfectly matched with Newtonian calculations. Indeed, analytically computing the future state of a system with more than two bodies and collisions may be impossible (Diacu, 1996), and even our best algorithms for physical simulation – computer physics engines – necessarily only approximate accurate Newtonian mechanics (Millington, 2010).

This is also not a claim that all of our physical intuitions are perfectly accurate, but rather that the approximations the mind makes when simulating everyday physical events are good enough to accomplish prediction and planning. Errors may still arise from approximations within our simulation engines: for instance, people often incorrectly judge the stability of asymmetric objects (Cholewiak, Fleming, & Singh, 2013) or how fast a wheel rim will roll down a slope (Proffitt et al., 1990), but these errors could be driven by simplifications in our representations of the shape of those objects (Ullman, Spelke, Battaglia, & Tenenbaum, 2017).

Instead we claim that for scenes with simple objects and physical principles that we encounter regularly (e.g., ballistic motion), our simulations will closely match the outcome of Newtonian equations. Using the terminology of Marr (1982), we suggest that people solve physical problems according to Newtonian mechanics at the *computational* level, but we do not make claims for how the mind does so at the *process* level. Studying exactly how people perform this physical simulation efficiently and mostly accurately is an exciting and open area for future research.

4.5 Conclusion

While casual observation and careful experimentation often suggest we exploit rich and accurate knowledge to interact with the world, the many errors and biases we make in other task regimes have driven important debates as to whether cognition is generally rational (Anderson, 1990; Tenenbaum, Kemp, Griffiths, & Goodman, 2011) or whether it is based on a set of ad hoc heuristics (Gigerenzer & Gaissmaier, 2011; Marcus & Davis, 2013). We find that in the domain of physical reasoning, there is a separate contrast: just as a basketball player might weave past opponents to score a basket but not be able to explain what he is about to do, or as we can all speak coherently without explicit knowledge of how to conjugate verbs, our ability to reason about physical events differs depending on how and why we are applying that knowledge. Thus the contrast between calibrated actions and error-prone reasoning is not just a result of having an approximately accurate understanding of some principles but not others, but rather because different domains of behavior rely on different cognitive facilities. Rather than argue whether people do or do not understand certain physical principles, we should therefore study the different systems people have for physical reasoning and how we choose to apply those systems across different tasks.

Acknowledgments

The authors would like to thank Frank Jäkel, Nancy Kanwisher, Joshua Tenenbaum, Gary Marcus, and Ernest Davis for their helpful discussion and comments, as well as Chaz Firestone and one anonymous reviewer for their insightful feedback.

KAS and EV were supported by NSF CPS grant 1239323, and a UCSD Interdisciplinary Collaboratories grant. KAS, EV and PB were supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of the Interior (DOI) contract D10PC20023. KAS was supported by CBMM funded by NSF STC award CCF-1231216, and ONR grant N00014-13-1-0333.

References

Aglioti, S., DeSouza, J. F. X., & Goodale, M. A. (1995). Size-contrast illusions deceive the eye but not the hand. *Current Biology, 5*, 679-685.

Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Erlbaum.

Bates, C., Battaglia, P., Yildirim, I., & Tenenbaum, J. B. (2015). *Humans predict liquid dynamics using probabilistic simulation.* Paper presented at the Proceedings of the 37th Annual Conference of the Cognitive Science Society.

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

Camerer, C. F. (1987). Do biases in probability judgments matter in markets? Experimental evidence. *The American Economic Review, 77*(5), 981-997.

Caramazza, A., McCloskey, M., & Green, B. (1981). Naive beliefs in "sophisticated" subjects: Misconceoptions about trajectories of objects. *Cognition, 9*, 117-123.

Chen, S. Y., Ross, B. H., & Murphy, G. L. (2014). Implicit and explicit processes in category-based induction: Is induction best when we don't think? *Journal of Experimental Psychology: General, 143*(1), 227-246.

Cholewiak, S. A., Fleming, R. W., & Singh, M. (2013). Visual perception of the physical stability of asymmetric three-dimensional objects. *Journal of Vision, 13*(12). doi:10.1167/13.4.12

Davis, E., Marcus, G., & Frazier-Logue, N. (2017). Commonsense reasoning about containers using radically incomplete information. *Artificial Intelligence, 248*, 46-84.

Dawson, M. R. W. (1988). Fitting the ex-Gaussian equation to reaction time distributions. *Behavior Research Methods, Instruments & Computers, 20*(1), 54-57.

Diacu, F. (1996). The solution of the n-body problem. *The Mathematical Intelligencer, 18*(3), 66-70.

diSessa, A. A. (1993). Toward an epistemology of physics. *Cognition and Instruction, 10*(2&3), 105-225.

Faisal, A. A., & Wolpert, D. M. (2009). Near optimal combination of sensory and motor uncertainty in time during a naturalistic perception-action task. *Journal of Neurophysiology, 101*, 1901-1912.

Fischer, J., Mikhael, J. G., Tenenbaum, J. B., & Kanwisher, N. (2016). Functional neuroanatomy of intuitive physical inference. *Proceedings of the National Academy of Sciences, 113*(34), E5072-E5081.

Frick, A., Huber, S., Reips, U.-D., & Krist, H. (2005). Task-specific knowledge of the law of pendulum motion in children and adults. *Swiss Journal of Psychology, 64*(2), 103-114.

Gerstenberg, T., Peterson, M. F., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2017). Eye-tracking causality. *Psychological Science, 28*(12), 1731-1744.

Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic decision making. *Annual Review of Psychology, 62*, 451-482.

Glaser, C., Trommershäuser, J., Mamassian, P., & Maloney, L. T. (2012). Comparison of the distortion of probability information in decision under risk and an equivalent visual task. *Psychological Science, 23*(4), 419-426.

Glover, S. (2004). Separate visual representations in the planning and control of action. *Behavioral and Brain Sciences, 27*, 3-78.

Goodale, M. A., & Milner, A. D. (1992). Separate visual pathways for perception and action. *Trends in Neurosciences, 15*(1), 20-25.

Griffiths, T. L., & Tenenbaum, J. B. (2006). Optimal prediction in everyday cognition. *Psychological Science, 17*(9), 767-773.

Hamrick, J. B., Smith, K. A., Griffiths, T. L., & Vul, E. (2015). *Think again? The amount of mental simulation tracks uncertainty in the outcome.* Paper presented at the 37th Annual Meeting of the Cognitive Science Society.

Hecht, H., & Bertamini, M. (2000). Understanding projectile acceleration. *Journal of Experimental Psychology: Human Perception and Performance, 26*(2), 730-746.

Hegarty, M. (2004). Mechanical reasoning by mental simulation. *Trends in Cognitive Sciences, 8*(6), 280-285. doi:10.1016/j.tics.2004.04.001

Hubbard, T. L. (1997). Target size and displacement along the axis of implied gravitational attraction: Effects of implied weight and evidence of representational gravity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23*(6), 1484-1493.

Kahneman, D. (2011). *Thinking, Fast and Slow*. New York, NY: Farrar, Straus and Giroux.

Kaiser, M. K., Proffitt, D. R., & Anderson, K. (1985). Judgments of natural and anomalous trajectories in the presence and absence of motion. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11*(4), 795.

Kaiser, M. K., Proffitt, D. R., & McCloskey, M. (1985). The development of beliefs about falling objects. *Attention, Perception, & Psychophysics, 38*(6), 533-539.

Kaiser, M. K., Proffitt, D. R., Whelan, S. M., & Hecht, H. (1992). Influence of animation on dynamical judgments. *Journal of Experimental Psychology: Human Perception and Performance, 18*(3), 669-689.

Kerzel, D. (2002). The locus of "memory displacement" is at least partially perceptual: Effects of velocity, expectation, friction, memory averaging, and weight. *Perception and Psychophysics, 64*(4), 680-692.

Kozhevnikov, M., & Hegarty, M. (2001). Impetus beliefs as default heuristics: Dissociation between explicit and implicit knowledge about motion. *Psychonomic Bulletin & Review, 8*(3), 439-453.

Kubricht, J. R., Jiang, C., Zhu, Y., Zhu, S.-C., Terzopoulos, D., & Lu, H. (2016). *Probabilistic simulation predicts human performance on viscous fluid-pouring problem.* Paper presented at the Proceedings of the 38th Annual Conference of the Cognitive Science Society.

Lieder, F., & Griffiths, T. L. (2017). Strategy selectino as rational metareasoning. *Psychological Review, 124*(6).

Marcus, G. F., & Davis, E. (2013). How robust are probabilistic models of higher-level cognition? *Psychological Science*. doi:10.1177/0956797613495418

Marr, D. (1982). *Vision*. Cambridge, MA: MIT Press.

McCloskey, M. (1983). Naive theories of motion *Mental models* (pp. 299-324).

McCloskey, M., Caramazza, A., & Green, B. (1980). Curvilinear motion in the absence of external forces: Naive beliefs about the motion of objects. *Science, 210*(5), 1139-1141.

McCloskey, M., & Kohl, D. (1983). Naive physics: The curvilinear impetus principle and its role in interactions with moving objects. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 9*(1), 146.

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.

McIntyre, J., Zago, M., Berthoz, A., & Lacquaniti, F. (2001). Does the brain model Newton's laws? *Nature Neuroscience, 4*(7), 693-694.

Millington, I. (2010). *Game physics engine development: How to build a robust commercial-grade physics engine for your game*. Boca Raton, FL: Taylor and Francis.

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1988). Adaptive strategy selection in decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14*(3), 534-552.

Piantadosi, S. T., Tily, H., & Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences, 108*(9), 3526-3529.

Pittenger, J. B. (1985). Estimation of pendulum length from information in motion. *Perception, 14*, 247-256.

Proffitt, D. R., & Gilden, D. L. (1989). Understanding natural dynamics. *Journal of Experimental Psychology: Human Perception and Performance, 15*(2), 384-393.

Proffitt, D. R., Kaiser, M. K., & Whelan, S. M. (1990). Understanding wheel dynamics. *Cognitive Psychology, 22*(3), 342-373.

Ranney, M. (1994). Relative consistency and subjects' "theories" in domains such as naive physics: Common research difficulties illustrated by Cooke and Breedin. *Memory & Cognition, 22*(4), 494-502.

Russell, S., & Wefald, E. (1991). Principles of metareasoning. *Artificial Intelligence, 49*(1-3), 361-395.

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.

Schwartz, D. L., & Black, J. B. (1996). Analog imagery in mental model reasoning: Depictive models. *Cognitive Psychology, 30*(2), 154-219.

Shanon, B. (1976). Aristotelianism, Newtonianism and the physics of the layman. *Perception, 5*(2), 241-243.

Smith, K. A., Battaglia, P., & Vul, E. (2013). *Consistent physics underlying ballistic motion prediction.* Paper presented at the 35th Annual Conference of the Cognitive Science Society, Berlin, Germany.

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

Stocker, A. A., & Simoncelli, E. P. (2006). Noise characteristics and prior expectations in human visual speed perception. *Nature Neuroscience, 9*(4), 578-585.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science, 331*(6022), 1279-1285.

Trommershäuser, J., Landy, M. S., & Maloney, L. T. (2006). Humans rapidly estimate expected gain in movement planning. *Psychological Science, 17*(11), 981-988.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review, 90*(4), 293-315.

Ullman, T. D., Spelke, E., Battaglia, P., & Tenenbaum, J. B. (2017). Mind Games: Game Engines as an Architecture for Intuitive Physics. *Trends in Cognitive Sciences, 21*(9), 649-665.

Wolpert, D. M., Ghahramani, Z., & Jordan, M. I. (1995). An internal model for sensorimotor integration. *Science, 269*(5232), 1880-1882.

Wu, S.-W., Delgado, M. R., & Maloney, L. T. (2009). Economic decision-making compared with an equivalent motor task. *Proceedings of the National Academy of Sciences, 106*(15), 6088-6093.

Zago, M., Bosco, G., Maffei, V., Iosa, M., Ivanenko, Y. P., & Lacquaniti, F. (2004). Internal models of target motion: Expected dynamics overrides measured kinematics in timing manual interceptions. *Journal of Neurophysiology, 91*, 1620-1634.

Zago, M., & Lacquaniti, F. (2005). Cognitive, perceptual and action-oriented representations of falling objects. *Neuropsychologia, 43*, 178-188.