Estimating Intuitive Physical Parameters Using Markov Chain Monte Carlo with People

Pat C Intara (pcl8798@nyu.edu)
Todd M Gureckis (todd.gureckis@nyu.edu)

Department of Psychology, New York University
6 Washington Place, New York, NY 10003

Abstract

A central question in cognitive science is the degree to which human and animal brains have adapted to and internalized the physical laws that govern the motion of objects. In this project, we propose a new method to estimate aspects of our intuitive sense of physical laws. Rather than assuming that humans internalize the form of Newtonian physics as found on Earth, we instead designed a procedure which allowed us to estimate which forms of physical laws feel most natural and intuitive to human participants. Our approach combines Markov chain Monte Carlo with People (MCMCP) and a custom parameterized physics engine. Each proposal of the MCMCP chain instantiated a world with new physical parameters and participants judged which of two scenes seemed more natural. Preliminary results show that this approach can quantify the precision of people's estimate of the direction and strength of gravity.

Keywords: intuitive physics; physical reasoning; cognitive models; Markov chain Monte Carlo

Introduction

The study of intuitive physics is concerned with people's everyday knowledge of how objects in the world move and interact. One key question has been the degree to which human and animal brains have adapted to and internalized the physical laws that govern the dynamics of objects. Under one influential account, humans possess a core reasoning engine which roughly implements Newtonian mechanics along with probabilistic noise. This theory, known as the *noisy Newton model* (Sanborn, Mansinghka, & Griffiths, 2009; Gerstenberg, Goodman, Lagnado, & Tenenbaum, 2012), holds that people have evolved or learned an accurate (if implicit) model of forces, collision dynamics, gravity, etc., but that errors in perceptual input and internal representation lead to the sometimes imperfect pattern of human performance (Smith & Vul, 2013; Sanborn, Mansinghka, & Griffiths, 2013; Kominsky et al., 2017). Even seemingly foundational errors, such as logical inconsistencies in trajectory estimation (Ludwin-Peery, Bramley, Davis, & Gureckis, 2021), can be accounted for with algorithmic simplifications and shortcuts in the mind's Newtonian physics engine (Ullman, Spelke, Battaglia, & Tenenbaum, 2017; Bass, Smith, Bonawitz, & Ullman, 2021).

However, Newtonian physics, and particularly the parametric instantiation of those rules as experienced on Earth (e.g., $F = ma$, $g \approx 9.81 m/s^2$), represent just one of the possible sets of physical laws. For example, people are familiar with video games that use unrealistic and artificial dynamics to make playing them more fun, and astronauts adapt quickly to life in microgravity (though not perfectly, see McIntyre, Zago, Berthoz, & Lacquaniti, 2001). Viewed more broadly, there is in fact a universe of possible alternative physics (a point of some concern to cosmology and philosophy, Carr & Rees, 1979; Barrow & Tipler, 1986). For example, while in our universe force F is equal to mass times acceleration (the second derivative of position), *psychologically* it is possible that it is equal to mass times velocity ($F = mv$) or that the gravitational constant g has a different value.

In this project, we propose a new method to estimate aspects of our intuitive sense of physical laws. Rather than assuming that humans internalize the ground-truth laws of Newtonian physics on Earth, we instead designed a procedure which allows us to *estimate* which form of physical laws feels more natural and intuitive to human participants. We did this by writing a custom physics engine in which we could "tinker" with the laws of physics. The custom physics engine can express a wide range of possible physics, many of which differ from the ones we experience on Earth in both parametric and functional form. We then elicited from participants judgments of which physical laws seemed most natural or correct. Our preliminary experiments with this method show promise, allowing us to estimate the "psychologically correct" form of various physical laws. After showcasing three experiments using the method, we lay out future directions and implications.

Estimating people's intuitive laws of physics using MCMC with People

In Markov chain Monte Carlo with People (MCMCP; introduced by Sanborn & Griffiths, 2007), participants repeatedly choose between two candidate stimuli that vary on some dimensions of interest. In a typical experiment, these might be two members of some category (e.g., *apple*) with the task of selecting which is a better example of that category (e.g., the rounder, redder one). Behind the scenes, these stimuli implement Markov chain Monte Carlo estimation: one of the options is the *current state* and the other is the *proposal*. Whichever is chosen becomes the state for the next trial. If the proposals are drawn from a suitably symmetric distribution—and with some reasonable assumptions about people's decision rules—these repeated selections serve as a valid acceptance function for ordinary Markov chain Monte

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

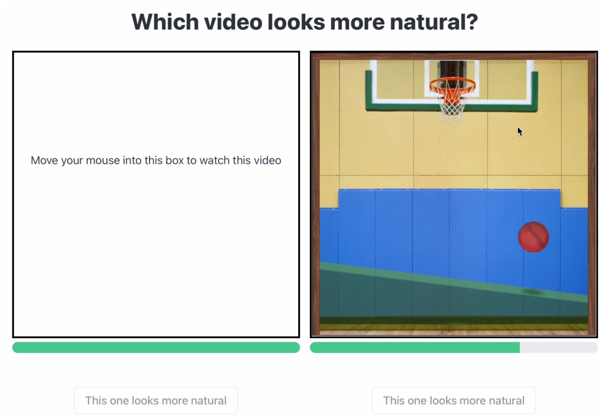

Figure 1: The task as seen by participants.

Carlo sampling. The values of the chain (after it has had time to settle) can therefore be interpreted as approximate samples of participants' internal representations of the relevant dimensions. This technique has been influential in studies of category representations (Hsu, Martin, Sanborn, & Griffiths, 2012; León-Villagrá, Otsubo, Lucas, & Buchsbaum, 2020), and, more relevant to the present work, has been used to evaluate people's perception of relative object mass in elastic collisions (Cohen & Ross, 2009).

In our experiments, the stimuli were animated videos of physical scenes with objects falling, bouncing, and colliding with each other, and participants selected which animation looked more natural. Although some research on intuitive physics has people reason about videos of real-life objects moving on Earth (Hood, 1995; Stahl & Feigenson, 2015; Little & Gureckis, 2024), it is increasingly popular to render videos of scenes using off the shelf computer graphics systems endowed with physics engines which are generally designed to be accurate (Todorov, Erez, & Tassa, 2012; Battaglia, Hamrick, & Tenenbaum, 2013; de Avila Belbute-Peres, Smith, Allen, Tenenbaum, & Kolter, 2018). Instead, in our work we developed a simple physics engine which was continuously parameterized in several ways (see below). As a result each proposal of our MCMC chain resulted in a new physics engine instantiation which was parameterized with different physical laws. Participants' choices then reflected their assessment of which physics seemed more "natural" to them between two alternatives. By randomizing the scene details (the initial positions and quantities of objects) our chains average over incidental aspects of any particular situation and expose which physics engine settings feel most natural, on average.

To validate our approach, our MCMC chains manipulated two physical constants: the direction of gravitational acceleration in Experiment 1, and the magnitude of gravitational acceleration in Experiments 2 and 3. In this way we could test a range of variables—including highly unrealistic ones—to estimate the shape of people's internal representations of these physical constants.

Experiment 1: Estimating gravity direction

On Earth, dropped objects fall directly downward[1], and people can reliably tell which direction that is, especially in built environments with many verticality cues (Haji-Khamneh & Harris, 2010). This made estimating the direction of gravity a straightforward validation of our procedure. We expected people to notice when the simulated gravity in our scenes did not point straight down, and the resulting distribution of natural gravity directions to be narrow and centered at zero degrees from vertical.

Method

Participants We recruited 50 English-speaking adult participants from Prolific and paid them each \$8.05 for participating in the experiment, which took approximately 32 minutes (for target pay rate of \$15/hr). Participants were excluded from our analyses if they did not complete the task and correctly answer 5 catch trials (which contrasted a normal bounce with a ball that slowly levitated while spinning rapidly). Participants were also excluded if they selected one of the videos more than 70% of the time or if their chains did not cross within the first 20 trials[2], leaving N=30.

Stimuli Participants watched animations of idealized balls falling and colliding with each other or with flat surfaces (see Figure 1). A demo of the task is available at `https://exps.gureckislab.org/e/roll-perpetual-sock`. Each trial had a 50% chance of featuring one ball and a 50% chance of two. The starting positions of the balls were sampled uniformly on each trial from a rectangular space above a ramp toward which they would fall and bounce (on trials with two balls, their positions were sampled from non-overlapping regions). The ramp angle was sampled uniformly on each trial from the range [-11.5°, 11.5°] from horizontal. The balls began each video with zero initial velocity but immediately began to fall downward—the motion of the balls was determined by a basic custom physics engine that used well-known simplifications for rigid body dynamics (Whittaker, 1904, pp. 231-232; Mirtich & Canny, 1995). This allowed the physical constants governing the animations to be set to arbitrary (and even impossible) values.

The starting positions of the balls and ramp were random on each trial but shared for the two animations on that trial. The only difference between the videos was the direction of gravity, which was controlled by the MCMC procedure. We

[1] Or at least very close to it: the presence of mountains does cause some variation in the local direction of gravity (known as *vertical deflection*), but it doesn't exceed 0.04° on Earth (Hirt et al., 2013).

[2] While sophisticated methods exist for evaluating the quality and convergence of MCMC chains (Vehtari, Gelman, Simpson, Carpenter, & Bürkner, 2021), we were limited in the length of chains that could practically be acquired with humans in the loop. A pilot experiment showed that the majority of participants' chains had crossed (a simple heuristic for establishing convergence; Sanborn & Griffiths, 2007) after 20 pairs of trials. For simplicity, we excluded any participants whose chains had not crossed within 20 pairs of trials, and excluded the first 20 pairs of trials as burn-in for all those who remained. This left 30 pairs of trials for each included participant.

used two alternating chains, so that the even numbered trials established the values for one chain and the odd numbered trials the other (this order was selected at random for each participant). On the first trial for each chain, the current state was one of the two initial chain values (-11.5° and 22.9°from vertical[3]), and the proposal was drawn from a wrapped normal distribution with mean equal to the current state and standard deviation of 9.2°. On each subsequent trial, the current state was the chosen proposal distribution was the same. For all trials, the left/right positions of the proposal and current state were decided at random.

The balls themselves were represented with a textured images meant to evoke larger sports balls like kickballs or basketballs with a diameter of approximately 25 cm. This styling was intended to provide a sense of scale, since previous work has shown that people more accurately estimate ballistic trajectories when they're displayed in a rich visual scene with cues to spatial scale (Miller et al., 2008). The other scene parameters (coefficient of restitution, moment of inertia, air drag, friction, etc.) were set to be compatible with the real-world properties of this kind of ball (Maynes, Compton, & Baker, 2005).

Task Participants watched the videos and selected the one they thought looked more natural (see Figure 1). Moving their cursor into either of the two square video frames would start the corresponding animation (which would stop when the cursor left the frame). While one animation played, the progress bar beneath it would fill, and when both bars were full the two buttons below would be enabled, allowing the participant to make their selection and advance to the next trial. Each progress bar took four seconds to fill, and the animations restarted after two seconds, so participants had to watch each animation at least twice. Although this enforced a minimum viewing time, subjects were free to look at each video for longer if they wanted.

Results and discussion

Figure 2 shows example chains of participant responses, and Figure 3 shows estimates of the resulting distributions. The distributions from the higher and lower chains are similar (means: -0.8 and 4.6, medians: -1.8 and 2.1, standard deviations: 13.1 and 14.0, 95% HDIs of means: [-19.9, 32.0] and [-23.0, 29.4]), and centered closer to the physically correct value (0°) than their starting values. Ideally, samples from MCMC chains should be independent of their starting values, which was not quite the case here (the values from the chain that started lower were slightly lower overall). Nevertheless, we considered the chains to have enough overlap to provide reasonable estimates of participants' internal representations, so for this and future experiments the values from both chains were taken together as a single distribution (here, mean: 1.92, median: 0.14, standard deviation 13.8, 95% HDI of mean: [-23.2, 29.2]).

[3]Or -0.2 and 0.4 radians. These values are asymmetric so that their average is different from zero, preventing participants from answering correctly by averaging the stimuli they see across the two chains.

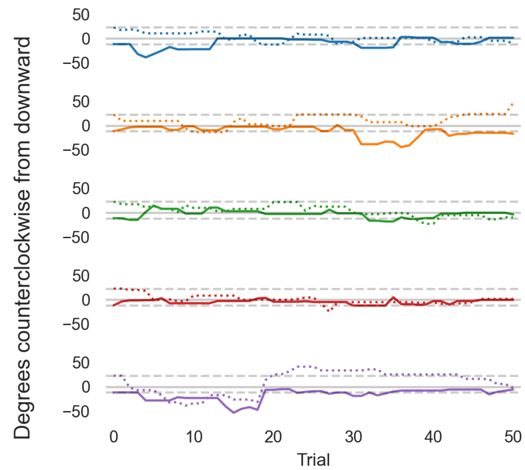

Figure 2: Lower chain values (solid line) and higher chain values (dotted line) from the first five included participants, along with the start points for each chain (lighter dashed lines), and physically correct value (lighter solid line).

The distributions in Figure 3 are tightly clustered around zero, as expected. One way to evaluate human performance is to consider each trial as having had a correct answer: whichever of the two stimuli (the current state or the proposal) had a gravity direction closer to directly downward. On this measure, all participants performed above chance (range: [56%, 85%], mean: 68.3%, median: 66%). Another way is consider an alternative where participants simply chose the video on the left or right with 50% probability. Since each trial would then have a 50% chance of accepting the proposal distribution, which itself is approximately normal, the chain values would be expected to follow a distribution with mean equal to the starting value and standard deviation σ_c equal to $\sigma_p\sqrt{N/2}$, where σ_p is the standard deviation of the proposal distribution and N is the number of trials in each chain. The expected value or the minimum and maximum value across both chains would then be σ_c below the lower chain starting value and σ_c above the higher chain starting value. To allow easy visual comparison of the empirical distributions across experiments and to the range expected by chance, the x-axis values for Figure 3 and Figure 4 are set to these minimum and maximum values.

These results serve to validate the use of MCMCP for estimating physical parameters—where chance behavior would have led to wide distributions, participants instead showed precise and accurate knowledge of the way things fall. Having confirmed that our method can produce reliable estimates for physical parameters that people know well, we can now evaluate behavior for parameters that people may represent less precisely.

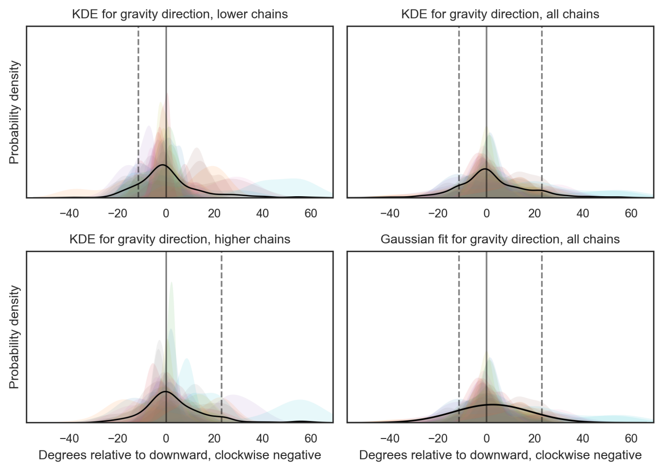

Figure 3: Kernel density estimates from the last 30 trials of all participants in Experiment 1 (each shaded region represents one participant's data), start points for each chain (dashed lines), physically correct value (solid line), along with aggregate summaries of all the included chain values: a kernel density estimate (upper row and lower left) and a Gaussian fit (lower right). The left column treats each participant's lower and higher chains separately, while the right column combines the two chains for each participant. For ease of comparison with Figure 4, all x-axis limits are the expected maximum and minimum chain values if participants had made their selections at random.

Experiment 2: Estimating gravity strength

There is a substantial literature on the question of whether people have a stable internal model of gravity, as we might use for everyday motor behaviors like catching thrown objects, with substantial evidence to support it (McIntyre et al., 2001; Zago et al., 2004; La Scaleia, Zago, & Lacquaniti, 2015; Jörges & López-Moliner, 2017), but also reasons to doubt its precision (Baurès, Benguigui, Amorim, & Siegler, 2007), including systematic biases in expected behavior of pendulums (Bozzi, 1958; Pittenger, 1990; Frick, Huber, Reips, & Krist, 2005) and falling objects (Vicovaro, Noventa, & Battaglini, 2019; Gravano, Zago, & Lacquaniti, 2017). Here, we applied our MCMCP approach to this question to test participants' sensitivity to unrealistic gravity.

Method

All methodological details for Experiment 2 were identical to those of Experiment 1 except where noted.

Participants We recruited another 50 adults from Prolific and applied the same exclusion criteria, which left N=35.

Stimuli Videos were the same as those used in Experiment 1 except that the direction of gravity was now fixed at 0° (directly downward) and instead the magnitude of gravitational acceleration varied between the two videos on each trial. The two chains were started at 4.3 and 14.9 meters per squared second (m/s^2), whereas gravitational acceleration on Earth is approximately 9.81 m/s^2. The standard deviation of the proposal distribution was 4.3 m/s^2.

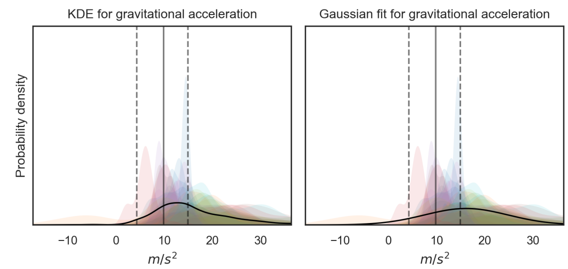

Figure 4: Kernel density estimates from the last 30 trials of all participants in Experiment 2 (shaded regions), start points for each chain (dashed lines), physically correct value (solid line), and summaries: kernel density estimate (left) and Gaussian fit (right). As in Figure 3, the x-axis limits are the expected minimum and maximum values if participants selected at random.

Results and discussion

The results of Experiment 2 are summarized in Figure 4. Pooled together, the chain values were in the vicinity of the correct value of 9.81 m/s^2, but with substantial variation (mean: 16.2, median: 14.5, standard deviation: 8.6, 95% HDI of mean: [4.2, 36.1]). Directly comparing the results of Experiments 1 and 2 is difficult because the units of Experiments 1 and 2 are different (it would not be valid, for example, to say that people are worse at estimating the magnitude of gravitational acceleration than its direction). Nevertheless, the relative widths of the distributions suggest that, in this experimental setting, participants' responses were closer to chance level when estimating the strength of gravity.

A major concern with strongly interpreting these results as an overestimation of the strength of gravity is that they could be equivalently explained as an underestimation of the depicted size of the scene[4]. A participant who responded in accordance with the overall median chain value of 14.5 m/s^2 could have thought gravity was 50% stronger than it really is, but could also have simply thought the ball was two-thirds as big. In the final experiment, we tested this more directly by manipulating the implied scale of the scene.

Experiment 3: Sensitivity to scale

Because modified gravitational acceleration is the same as normal gravity with the scale of the scene changed, in Experiment 3 we sought to distinguish inaccurate judgments of

[4]For example, a marble and a basketball both fall approximately 5 cm in the first 100 ms after being dropped, but for the marble this is three times its diameter and for a basketball it's barely one-fifth. If the marble and basketball are the same size on a video screen, say 100 pixels, then the marble falls 300 px in the same time that the basketball falls 20 px. Therefore, the marble, being 1/15th the size, accelerates 15 times faster (in pixel terms).

Figure 5: The scenes used in Experiment 3: *Gym* (left) and *Desk* (right).

natural physical motion from merely imprecise judgments of scene scale by directly asking participants to report the size of the ball. We also sought to determine whether participants' judgments were well-calibrated to the implied scale of the scene by having two different scene conditions.

Method

All methodological details for Experiment 3 were identical to those of Experiments 1 and 2 except where noted.

Participants We recruited another 40 adults from Prolific and applied the same exclusion criteria, leaving N=29.

Stimuli To make the scale and relative position of the ball easier to interpret, the ramp used in Experiments 1 and 2 was replaced with a slanted table, and more details were added to the background (see Figure 5). Half of participants were assigned at random to the *gym* condition, where the videos depicted a large ball bouncing in a gym as before (Figure 5, left) and half were assigned to the *desk* condition (Figure 5, right), where the videos depicted a small ball bouncing on a desktop. Crucially, the balls were the same size *in pixels* (radius of 22 px)—that is, regardless of condition, the ball took up the same amount of space on the screen.

Because participants' chain values would be scaled by their size judgments (see following sections), their adjusted starting chain values would also differ. Therefore, we sampled each of the two chain starts uniformly between 0 and 3.46 pixels/frame^2. Note the use of pixel units rather than absolute distance—this is because physically correct motion in the *desk* condition involves the ball accelerating faster across participants' screens, and therefore the two conditions have different correct answers in pixel units. To accommodate this faster apparent motion of the *desk* condition, the physics engine was modified to more precisely track ball movement between frames and avoid intersecting surfaces.

Task

In addition to selecting the more natural video, participants also reported the size of the ball using a slider with 1/4-inch increments. Other familiar sports balls (e.g., a golf ball, a tennis ball, etc.) were depicted to scale and with their dimensions visible so participants would have a clear reference when making their size judgments.

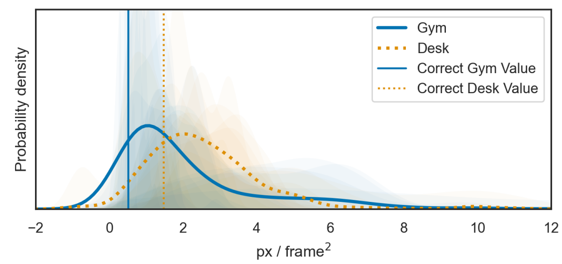

Figure 6: Kernel density estimates from the last 30 trials of all participants of Experiment 3 (shaded regions, colored by condition), correct answers for each condition (vertical lines), physically correct value (solid line), and summary kernel density estimate for each condition. Values are scaled to account for differences in inferred scene scale (see Results and discussion).

Results and discussion

We first used the mean of the size judgments for each condition to determine the implied scale and correct gravitational acceleration for each condition. Because participants varied in their judgments of the size of the scene, we scaled each participant's responses to account for the difference between their judgment and the mean value for their condition. For example, the mean ball size for the gym condition was 9.25 inches. If a given participant thought the ball was half that size, they would presumably have preferred videos where the gravity (in pixel units) was twice as strong, so their chain values would all be multiplied by one-half to account for that. This allowed us to fairly evaluate the accuracy of participants' gravity judgments, as distinct from any limitations in inferring the scale of the scene.

These adjusted chain values are summarized in Figure 6. The qualitative pattern from Experiment 2 was repeated. For the *gym* condition, where the correct value was 0.51 pixels/frame^2, the pooled chain values had mean: 2.30, median: 1.43, standard deviation: 2.5, 95% HDI of mean: [0.13, 6.75]. For the *desk* condition, where the correct value was 1.47 pixels/frame^2, the pooled chain values had mean: 2.60, median: 2.36, standard deviation: 1.7, 95% HDI of mean: [0.43, 5.66]. Chain values were larger for the *desk* condition on average (95% HDI of mean difference, *desk* minus *gym*: [0.29, 0.33]), as accurate physics estimation would require, suggesting that participants' naturalness judgments, while not very precise, were at least sensitive to the scene scale manipulation.

It is worth noting that our results here differ from those of other experiments (e.g., Vicovaro et al., 2019; Gravano et al., 2017) finding that people underestimate the strength of gravity, i.e., they expect things to fall more slowly than they do in reality (at least for small objects; Bozzi, 1958). Participants in this task were, if anything, more likely to overestimate (though not robustly: the 95% HDIs for both conditions included a range of values below the correct gravitational accelerations).

General discussion

In this report, we describe a novel method for estimating people's intuitive physical laws. Our approach combines MCMC with People (MCMCP) and a custom physics engine which can instantiate a wide range of physical laws. We conducted three experiments assessing the viability of this approach, including estimating both obvious and slightly less obvious laws and properties. We view these experiments as preliminary demonstrations of what can be done with this method. In future work we hope to explore even more alien regimes of the universe of possible physics.

One unique feature of our experiments is that perturbations to the underlying physics made on each trial were small, and occurred while relatively large changes to the scene details occurred across successive trials (the number of balls, their initial positions, and the angle of the ramp). Subjectively, this had the effect of weakening the utility of any single heuristic that could be used on every trial, potentially forcing participants to form a more general impression of the videos' overall physical naturalness. Future experiments will investigate whether participants indeed are less likely to rely on scene-specific heuristics—if so, this method would be well suited to elicit intuitive judgments about general physical laws quickly and efficiently.

One important question is how important the MCMCP procedure is for our conclusions. For example, another approach could be to allow participants to watch a single video playing in a loop and adjust a slider to change the underlying physics until they seem natural (Harrison et al., 2020). We aim to run such a comparison in the future, but we also believe the MCMCP procedure has some distinct advantages, including providing implicit estimates of the uncertainty in the target distribution (which may reflect some combination of perceptual limitations and cognitive uncertainty). These inferred distributions can then be reused in computational models. For example, models that include estimates of error in predicting object motion and interactions (e.g., Gerstenberg, Goodman, Lagnado, & Tenenbaum, 2021) could be built on more precise foundations. Even animators and video game designers might benefit from a better understanding of the acceptable range of imprecision when creating complex and computationally expensive animations (O'Sullivan, Dingliana, Giang, & Kaiser, 2003; Yeh, Reinman, Patel, & Faloutsos, 2009).

Finally, our results primarily assess physical laws through observation. However, different response paradigms might probe different internal representations (Zago & Lacquaniti, 2005; Smith, Battaglia, & Vul, 2018). For example, catching a ball is thought to tap into different mental resources than judging the naturalness of its trajectory, which itself may rely on different processes than explicitly drawing the shape of its arc (Kozhevnikov & Hegarty, 2001; Smith, Battaglia, & Vul, 2013). Our approach, therefore, can only reveal the areas of unnatural physics that seem natural when observed (Kaiser, Proffitt, & Anderson, 1985), not those that *feel* natural to physically experience (Won, Gross, & Firestone, 2021) or are intuitive when reasoned about explicitly (McCloskey, Caramazza, & Green, 1980). However, future work might be able to extend this to limited forms of active control (Bramley, Gerstenberg, Tenenbaum, & Gureckis, 2018).

Acknowledgments

The authors would like to thank members of the NYU Computation and Cognition Laboratory and the anonymous reviewers for their valuable feedback. Support provided by NSF BCS-2121102 awarded to T.M.G.

References

Barrow, J. D., & Tipler, F. J. (1986). *The anthropic cosmological principle.* Oxford University Press.

Bass, I., Smith, K. A., Bonawitz, E., & Ullman, T. D. (2021). Partial mental simulation explains fallacies in physical reasoning. *Cognitive Neuropsychology*, *38*, 413–424.

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*, 18327–18332.

Baurès, R., Benguigui, N., Amorim, M.-A., & Siegler, I. A. (2007). Intercepting free falling objects: better use occam's razor than internalize newton's law. *Vision research*, *47*(23), 2982–2991.

Bozzi, P. (1958). Analisi fenomenologica del moto pendolare armonico [phenomenological analysis of pendular harmonic motion]. *Rivista di Psicologia*, *52*(4), 281–302.

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

Carr, B. J., & Rees, M. J. (1979). The anthropic principle and the structure of the physical world. *Nature*, *278*(5705), 605–612.

Cohen, A. L., & Ross, M. G. (2009). Exploring mass perception with markov chain monte carlo. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(6), 1833.

de Avila Belbute-Peres, F., Smith, K., Allen, K., Tenenbaum, J., & Kolter, J. Z. (2018). End-to-end differentiable physics for learning and control. *Advances in neural information processing systems*, *31*.

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., Goodman, N., Lagnado, D., & Tenenbaum, J. (2012). Noisy newtons: Unifying process and dependency accounts of causal attribution. In *Proceedings of the annual meeting of the cognitive science society.*

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*(5), 936.

Gravano, S., Zago, M., & Lacquaniti, F. (2017). Mental imagery of gravitational motion. *Cortex*, *95*, 172–191.

Haji-Khamneh, B., & Harris, L. R. (2010). How different types of scenes affect the subjective visual vertical (svv) and the perceptual upright (pu). *Vision Research*, *50*(17), 1720–1727.

Harrison, P., Marjieh, R., Adolfi, F., van Rijn, P., Anglada-Tort, M., Tchernichovski, O., … Jacoby, N. (2020). Gibbs sampling with people. *Advances in neural information processing systems*, *33*, 10659–10671.

Hirt, C., Claessens, S., Fecher, T., Kuhn, M., Pail, R., & Rexer, M. (2013). New ultrahigh-resolution picture of earth's gravity field. *Geophysical research letters*, *40*(16), 4279–4283.

Hood, B. M. (1995). Gravity rules for 2-to 4-year olds? *Cognitive Development*, *10*(4), 577–598.

Hsu, A., Martin, J., Sanborn, A., & Griffiths, T. (2012). Identifying representations of categories of discrete items using markov chain monte carlo with people. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 34).

Jörges, B., & López-Moliner, J. (2017). Gravity as a strong prior: implications for perception and action. *Frontiers in Human Neuroscience*, *11*, 259425.

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.

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.

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

La Scaleia, B., Zago, M., & Lacquaniti, F. (2015). Hand interception of occluded motion in humans: a test of model-based vs. on-line control. *Journal of Neurophysiology*, *114*(3), 1577–1592.

León-Villagrá, P., Otsubo, K., Lucas, C., & Buchsbaum, D. (2020). Uncovering category representations with linked mcmc with people. In *Proceedings of the annual meeting of the cognitive science society.*

Little, P., & Gureckis, T. (2024). A nonparametric model of object discovery. In *Proceedings of the annual meeting of the cognitive science society.*

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

Maynes, K., Compton, M., & Baker, B. (2005). Coefficient of restitution measurements for sport balls: an investigative approach. *The Physics Teacher*, *43*(6), 352–354.

McCloskey, M., Caramazza, A., & Green, B. (1980). Curvilinear motion in the absence of external forces: Naive beliefs about the motion of objects. *Science*, *210*(4474), 1139–1141.

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

Miller, W. L., Maffei, V., Bosco, G., Iosa, M., Zago, M., Macaluso, E., & Lacquaniti, F. (2008). Vestibular nuclei and cerebellum put visual gravitational motion in context. *Journal of neurophysiology*, *99*(4), 1969–1982.

Mirtich, B., & Canny, J. (1995). Impulse-based simulation of rigid bodies. In *Proceedings of the 1995 symposium on interactive 3d graphics* (pp. 181–188).

O'Sullivan, C., Dingliana, J., Giang, T., & Kaiser, M. K. (2003). Evaluating the visual fidelity of physically based animations. *ACM Transactions on Graphics (TOG)*, *22*(3), 527–536.

Pittenger, J. B. (1990). Detection of violations of the law of pendulum motion: Observers' sensitivity to the relation between period and length. *Ecological Psychology*, *2*(1), 55–81.

Sanborn, A. N., & Griffiths, T. (2007). Markov chain monte carlo with people. *Advances in neural information processing systems*, *20*.

Sanborn, A. N., Mansinghka, V., & Griffiths, T. L. (2009). A bayesian framework for modeling intuitive dynamics. In *Proceedings of the annual meeting of the cognitive science society.*

Sanborn, A. N., Mansinghka, V. K., & Griffiths, T. L. (2013). Reconciling intuitive physics and newtonian mechanics for colliding objects. *Psychological review*, *120*(2), 411.

Smith, K. A., Battaglia, P., & Vul, E. (2013). Consistent physics underlying ballistic motion prediction. *Proceedings of the annual meeting of the cognitive science society*, *35*.

Smith, K. A., Battaglia, P. W., & Vul, E. (2018). Different physical intuitions exist between tasks, not domains. *Computational Brain & Behavior*, *1*, 101–118.

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

Stahl, A. E., & Feigenson, L. (2015). Observing the unexpected enhances infants' learning and exploration. *Science*, *348*(6230), 91–94.

Todorov, E., Erez, T., & Tassa, Y. (2012). Mujoco: A physics engine for model-based control. In *2012 ieee/rsj international conference on intelligent robots and systems* (pp. 5026–5033).

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.

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, P.-C. (2021). Rank-normalization, folding, and localization: An improved $\hat{R}$ for assessing convergence of mcmc (with discussion). *Bayesian analysis*, *16*(2), 667–

718.

Vicovaro, M., Noventa, S., & Battaglini, L. (2019). Intuitive physics of gravitational motion as shown by perceptual judgment and prediction-motion tasks. *Acta psychologica*, *194*, 51–62.

Whittaker, E. T. (1904). *A treatise on the analytical dynamics of particles and rigid bodies*. Cambridge University Press.

Won, I., Gross, S., & Firestone, C. (2021). "Impossible" somatosensation and the (ir)rationality of perception. *Open Mind*, *5*, 30–41.

Yeh, T. Y., Reinman, G., Patel, S. J., & Faloutsos, P. (2009). Fool me twice: Exploring and exploiting error tolerance in physics-based animation. *ACM Transactions on Graphics (TOG)*, *29*(1), 1–11.

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*(4), 1620–1634.

Zago, M., & Lacquaniti, F. (2005). Cognitive, perceptual and action-oriented representations of falling objects. *Neuropsychologia*, *43*(2), 178–188.