A simulation-heuristics dual-process model for intuitive physics

Shiqian Li[★,1,2]
shiqianli@stu.pku.edu.cn

Yuxi Ma[★,1]
mayuxi@bigai.ai

Jiajun Yan[★,1]
yjejuy@gmail.com

Bo Dai[2]
daibo@bigai.ai

Yujia Peng[1,2,3,✉]
yujia_peng@pku.edu.cn

Chi Zhang[2,✉]
zhangchi@bigai.ai

Yixin Zhu[1,✉]
yixin.zhu@pku.edu.cn

★ equal contributors   ✉ corresponding authors

[1] Institute for Artificial Intelligence, Peking University   [2] State Key Laboratory of General Artificial Intelligence, BIGAI

[3] School of Psychological and Cognitive Sciences and Beijing Key Laboratory of Behavior and Mental Health, Peking University

https://github.com/lishiqianhugh/DualModel

Abstract

The role of mental simulation in human physical reasoning is widely acknowledged, but whether it is employed across scenarios with varying simulation costs and where its boundary lies remains unclear. Using a pouring-marble task, our human study revealed two distinct error patterns when predicting pouring angles, differentiated by simulation time. While mental simulation accurately captured human judgments in simpler scenarios, a linear heuristic model better matched human predictions when simulation time exceeded a certain boundary. Motivated by these observations, we propose a dual-process framework, Simulation-Heuristics Model (SHM), where intuitive physics employs simulation for short-time simulation but switches to heuristics when simulation becomes costly. By integrating computational methods previously viewed as separate into a unified model, SHM quantitatively captures their switching mechanism. The SHM aligns more precisely with human behavior and demonstrates consistent predictive performance across diverse scenarios, advancing our understanding of the adaptive nature of intuitive physical reasoning.

Keywords: intuitive physics; physical reasoning; mental simulation; heuristic model

Introduction

Humans demonstrate extraordinary abilities in understanding and reasoning about the physical world even without formal training in physics (Piloto et al., 2022). This ability, known as intuitive physics (J. R. Kubricht et al., 2017), enables comprehending physical concepts (Baillargeon et al., 1985; Baillargeon & Graber, 1987; Kim & Spelke, 1992), predicting physical dynamics (P. W. Battaglia et al., 2013; C. Bates et al., 2015; Davis et al., 2017), and interacting with the physical environments (Allen et al., 2020). However, human intuitive physics may exhibit errors and biases in certain physical scenarios, indicating deviations from classical Newtonian physics (McCloskey et al., 1980, 1983; Kaiser et al., 1986; Kozhevnikov & Hegarty, 2001). Such errors and biases serve as a unique aspect of human reasoning, offering a valuable avenue for studying the underlying mechanisms of intuitive physics (J. R. Kubricht et al., 2017).

A common perspective to understanding human intuitive physics is mental simulation: it hypothesizes an approximate intuitive physics engine in the human mind (P. W. Battaglia et al., 2013; K. A. Smith & Vul, 2013; T. D. Ullman et al., 2017; K. Smith et al., 2024). This simulation framework, grounded in probabilistic inference, was found to be able to characterize human behavior across various physical tasks, and also account for human errors and biases, further validating its relevance and applicability (P. Battaglia et al., 2012; J. Kubricht

et al., 2016, 2017; Gerstenberg et al., 2017; T. D. Ullman et al., 2018; C. J. Bates et al., 2019; Bass et al., 2021; Chen et al., 2023; Li et al., 2023). Nevertheless, the simulation model fails to completely explain the variance in human behavior in some demanding or unfamiliar conditions (Schwartz & Black, 1999; Kozhevnikov & Hegarty, 2001; K. A. Smith et al., 2018; Ludwin-Peery et al., 2020, 2021), suggesting the existence of alternative cognitive mechanisms, possibly mental shortcuts employed for certain physical scenarios, or heuristics (Davis, 1998; Davis et al., 2017; Kozhevnikov & Hegarty, 2001; J. R. Kubricht et al., 2017; K. A. Smith et al., 2018). Recent work has further explored these limitations, showing that mental simulation has severe capacity constraints. Balaban & Ullman (2024) found that people can track only a single moving object in imagination. Li et al. (2023) showed that people use simplified object approximations for physical reasoning, employing shortcuts that reduce computational demands. Particularly relevant to our work, T. Ullman & Wang (2023) identified a switch point in a fluid-reasoning task where prediction patterns changed based on simulation demands. These findings in capacity constraints raise important questions about how humans might adapt their reasoning strategies when faced with increasingly complex physical scenarios.

Here we ask the questions: *Do humans consistently rely on mental simulation, or do they employ alternative heuristic strategies under certain conditions? What are the situations that prompt a switch between these two cognitive strategies?*

Previous studies have investigated the interplay between simulation and heuristics, providing evidence for **qualitative** insights. For instance, Kozhevnikov & Hegarty (2001) demonstrates that people tend to use impetus heuristics in quick judgment scenarios, while P. W. Battaglia et al. (2013) finds that models based on height heuristics can more accurately explain human judgment in certain tasks, such as predicting the falling distance of a block tower. Furthermore, K. A. Smith et al. (2017) and Yildirim et al. (2017) suggest that humans integrate these two cognitive strategies in motion prediction and physical planning tasks, respectively. However, there is currently no study that has provided clear evidence supporting the relationship between these two strategies or **quantitatively** demonstrated the transition between them. A comprehensive exploration is needed to understand whether a switch of policies exists and, if so, how these switches operate, as well as to identify alternative heuristics that could reverse engineer the human physical reasoning process, including human biases.

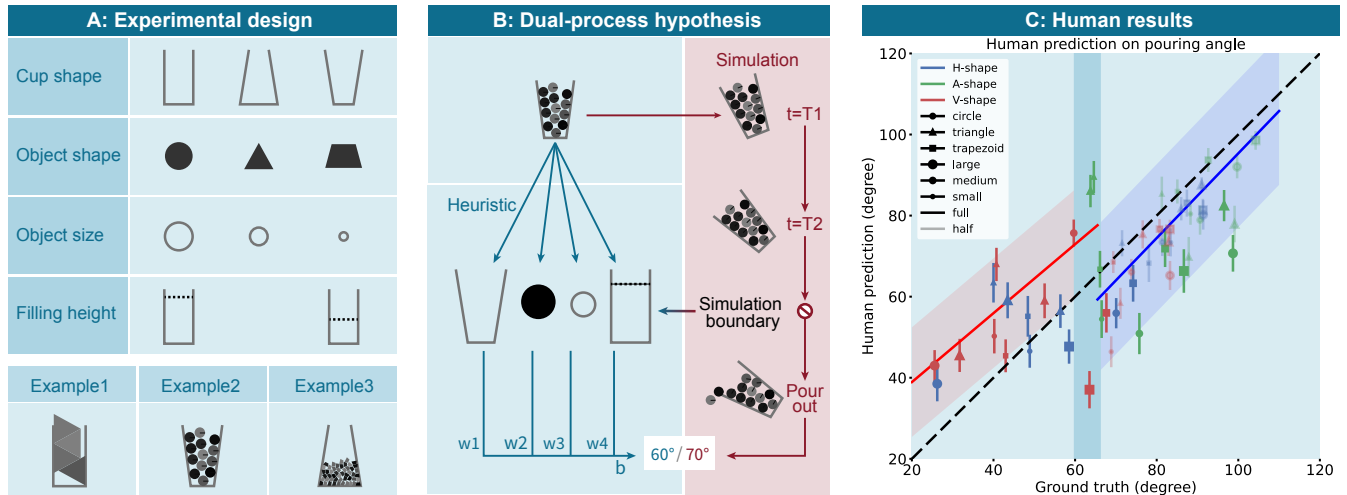## Judge the tilting angle at which the cup begins to pour out marbles

Figure 1: **(A) Experimental design:** Trials involved 3 cup shapes (H-shape, A-shape, V-shape), 3 object shapes (circle, triangle, trapezoid), 3 sizes (large, medium, small), and 2 filling heights (full, half), totaling 54 unique conditions. Participants predicted the tilt angle for marbles to fall out when cups are tilted to the left. **(B) SHM hypothesis:** Participants used either mental simulation, simulating the tilting process until pouring out, or a heuristic strategy, reaching judgments from physical features when the simulation exceeds a boundary. These methods could result in different outcomes. **(C) Human results:** Each point represents a condition, illustrating human tendencies to either overestimate or underestimate the pouring angle. The red and blue lines are the regression results of Intuitive Physics Engine (IPE) and the heuristic model, respectively. The SHM effectively captures human behavior with a switching boundary.

In our study, we systematically investigate the switch between simulation and heuristic strategies in intuitive physics, developing a computational model that offers improved explanatory power. We hypothesize that: (i) the simulation strategy prevails in scenarios simple enough for reliable physical unfolding; (ii) the heuristic strategy takes over when mental simulation becomes too costly; (iii) the switching point of the two strategies correlates with the simulation cost, approximated via a proxy of simulation time. Diverging from previous studies that often focused on simpler dynamics or predictable outcomes (P. W. Battaglia et al., 2013; K. A. Smith & Vul, 2013; K. A. Smith et al., 2017), our study engages in examining human reasoning across a range of simulation costs (Schwartz & Black, 1999; J. Kubricht et al., 2016; Davis et al., 2017). Inspired by previous pouring tasks in intuitive physics (Schwartz & Black, 1999; J. Kubricht et al., 2016; Guevara et al., 2017; Lopez-Guevara et al., 2020), we build a pouring marble task with more diverse physical properties and complexities. In this task, human participants are asked to judge the tilt angle needed to pour marbles from cups under various setups (see Fig. 1A).

We conducted four steps of experiment to validate the above hypotheses sequentially. The **first step** examines whether there is a pattern switch regarding human judgment. A finding of two distinct error patterns (*i.e.* overestimation and underestimation) supports the existence of two predominant strategies that vary under different simulation times. The **second step** aims to test our hypothesis on whether the IPE model can account for human judgments in simpler scenarios. The results show that it aligns well with human judgments and exhibits the same overestimation when the actual pouring angle is

small. However, it fails to account for humans' underestimation as the actual pouring angle exceeds a certain boundary (see Fig. 1C). Given that the pouring rate remains consistent, we hypothesize longer simulation time leads to increased cost of physical unfolding, triggering the transition to another cognitive strategy. Thus, we validate our second hypothesis by exploring an alternative heuristic approach in the **third step**. We developed a linear heuristic model trained on ground-truth data and found that, although less effective than IPE at smaller angles, the model accurately captures the underestimation pattern when the pouring angle exceeds a certain boundary. These results support our hypothesis of a cognitive shift to a heuristic strategy. To test our third hypothesis, in the **fourth step**, we explore whether a novel framework, Simulation-Heuristics Model (SHM), that combines these two models and toggles based on simulation cost, can explain human judgments across all complexity levels (see Fig. 1B). The results show that SHM aligns more closely with human behavior across diverse scenarios and metrics, enhancing our understanding of intuitive physical reasoning and highlighting the adaptability and versatility of human cognition.

Models

Mental simulation

Recent work explains human intuitive physics understanding by assuming an approximate simulation engine in the human mind (P. W. Battaglia et al., 2013; Lake et al., 2017; J. Kubricht et al., 2016). This engine serves to simulate the future physical unfolding, akin to a computational physics engine but incorporates noise into the physical properties of objects.

Following this approach, our model utilizes an IPE that runs

noisy simulations as in P. W. Battaglia et al. (2013). The model takes an initial physical scene S_0 and external forces $f_{0:T-1}$ to derive the judgment J. This process involves predicting the intermediate states $S_{1:T}$ over a time span T:

$$P(J|S_0, f_{0:T-1}) = \int_{S_{1:T}} P(J|S_{1:T})P(S_{1:T}|S_0, f_{0:T-1})dS_{1:T},$$
(1)

where $S_{1:T}$ represents the sequence of all physical states from time steps 1 through T, and the integral averages over all possible trajectories of these states. Each state evolves as $S_{t+1} = \phi(S_t + \epsilon, f_t)$ with noise $\epsilon \sim \mathcal{N}(0, \sigma^2)$, and $\phi(\cdot)$ representing deterministic physical dynamics. We simplify the mapping from the initial state to the final judgment as $M(S_0; f, \epsilon)$.

In our implementation using the flexible physics engine Pymunk, the IPE utilizes all physical variables to simulate future dynamics in a 2D scenario with added Gaussian noise $\mathcal{N}(0, \sigma^2)$ to each marble's position horizontally and vertically during every simulation step. When this noise causes objects to overlap, Pymunk's collision resolution automatically repels them based on physical constraints. The noise level σ^2 is varied from 0.1 to 1 to observe its impact on the simulation results. Inspired by classical findings by Shepard & Metzler (1971), who demonstrated that mental rotation time is proportional to rotation angle, suggesting a constant angular velocity in mental simulation. We assume a simulation model with constant rotational speed and optimize it through iterative parameter search to minimize the RMSE between model predictions and human judgments across a calibration subset of trials. After optimization, the simulation time becomes proportional to the rotation angle. *In the following context, we will use rotation angles to represent simulation time.*

We perform 30 noisy IPE simulations per trial. During each simulation, an automatic detection system is integrated to identify the moment when the marbles fall out, which serves as the ground truth. The final pouring angle is determined from the average of the 30 results. This setup allows us to mimic the variability and uncertainty in human cognition, as outlined in prior studies (K. A. Smith & Vul, 2013), and to explore how these factors influence judgment in physical tasks.

Heuristic model

Prior studies often employ predefined heuristics to elucidate human biases (Schwartz & Black, 1999; Kozhevnikov & Hegarty, 2001; K. A. Smith et al., 2017) or fit heuristic models on human data to evaluate the influence of physical attributes (Zhou et al., 2023). While these approaches offer insights for specific tasks, a systematic methodology for learning heuristics in complex scenarios is lacking.

Our heuristic model is designed to learn from a subset of physical attributes, fitting ground-truth data through a direct mapping g from the initial scene S_0 to the final judgment J, bypassing the intermediate states. This model is advantageous as it approximates humans' real-world physics understanding by a limited set of attributes, and circumvents the need for computationally heavy physics simulation. In particular, we employ a linear model with learnable parameters:

$$J = g(S_0^1, ..., S_0^n) = \sum_{i=1}^{n} \omega_i S_0^i + b,$$
(2)

where $\{S_0^i\}$ are different physical variables in S_0 and we set $n = 4$ in our study. Specifically, the model considers the following four variables: object size, filling height, object shape, and cup shape. Instead of directly predicting the pouring angle, the model predicts the difference between the actual pouring angle and a reference 90-degree angle. This design choice was made based on preliminary observation from a familiarization experiment that an H-shape cup containing little marbles almost always pours out at 90 degrees. The model is optimized using the mean squared error. Future exploration may consider nonlinear heuristic models using symbolic regression (Xu et al., 2021).

Dual-process model

Building on the notion that human cognition might employ multiple systems (Kahneman, 2011), we introduce a dual-process model in the context of intuitive physics, termed SHM. This model hypothesizes that humans alternate between two strategies, mental simulation and heuristic reasoning, based on the simulation cost, which is approximated by the simulation time. The simulation strategy parallels Kahneman's "slow thinking" (System 2): a deliberate, resource-intensive process, while the heuristic approach functions as "fast thinking" (System 1): an efficient but potentially less accurate mechanism. Specifically, for simulation time below a critical boundary θ, IPE is favored, whereas, beyond θ, a heuristic strategy is triggered. This adaptive approach is formalized as:

$$\begin{cases} J = E_\epsilon[M(S_0; \epsilon)], & \text{if } T \leqslant \theta, \\ J = \sum_{i=1}^{n} \omega_i S_0^i + b, & \text{if } T > \theta \end{cases}.$$
(3)

where we drop the dependency on f, which remains constant across the same set of experiments. We employ a grid search method to optimize both θ for the strategic transition and the noise parameters σ for the IPE, in addition to a group of heuristic parameters ω derived from linear regression.

Experiment

Participants

A total of 43 college students (55% male, 45% female; mean age = 21.77 ± 4.45) completed in-person experiments for course credit or monetary compensation. One participant was excluded due to little response variation. The study received IRB approval from the Committee for Ethics and the Protection of Human and Animal Welfare at the local institution.

Stimuli

Stimuli were generated using Pymunk in various configurations, encompassing three cup shapes (H-shape, A-shape, and V-shape), three object shapes (circle, triangle, and trapezoid), three object sizes (large, medium, and small), and two filling

heights (half and full), totaling up to 54 different conditions. These stimuli were then rendered using Pygame.

Marbles were randomly placed inside cups, with layouts adjusted by Pymunk's physics engine. Three random layouts were generated per condition. Stimuli were selected when marbles reached stability. Marbles had no friction or elasticity, equal mass, and random grayscale colors to eliminate material property assumptions.

Pouring angles were determined through simulations of slow cup rotations. The critical angle was measured when a marble's mass center aligned with the cup's top-left corner, using 120 FPS calculations with automatic fall detection. Tilt angles were referenced to verify pouring angles. Example stimuli appear in Fig. 2a.

Procedures

A within-subjects design was implemented, where each participant completed all 54 conditions. Stimuli were navigated in a counter-balanced order with randomly selected layouts, and the experiment lasted approximately 30 minutes.

Familiarization Following informed consent, participants underwent an instruction phase and familiarization session featuring pouring demonstrations with two small marbles, addressing marble properties, tilt angles, and pour initiation criteria. Comprehension was assessed through quizzes on empty cup angles and marble pour timing before proceeding to the experimental phase.
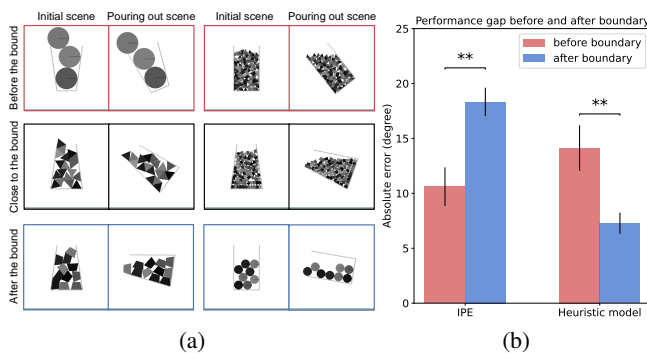

(a)   (b)

Figure 2: Visualizations of stimuli and error analysis. (a) Example stimuli. The top (red), middle (black), and bottom (blue) rows depict two scenarios each, with pouring angles that are smaller, close to, and larger than the established simulation bound, respectively. (b) The mean absolute error between model and human results (with SEM). The IPE model exhibits a larger absolute error when the simulation time exceeds the boundary. Conversely, the heuristic model shows contrary results, indicating its effectiveness in these scenarios.

Experiment Participants were required to complete 54 trials consecutively. In each trial, a static image of a non-rotated cup from various setups was presented. The tilt angle necessary for the cup to begin pouring out marbles was estimated by the participants using a slider bar with a range of 0 to 135 degrees. To reduce potential biases from inaccurate angle perception, a dial marked with angle measurements was provided in each trial. Demographic information along with the responses for the pouring angles across all 54 trials, including

the total duration, were recorded for subsequent analysis.

Results

Our analysis follows four sequential steps to validate our hypothesis: (1) examination of participants' error patterns indicating strategic shifts, (2) evaluation of IPE's capacity to explain human judgment, (3) development of a complementary heuristic model incorporating physical attributes, and (4) integration of simulation and heuristic approaches into a hybrid model to comprehensively explain human judgments.

A switching in error patterns

Human results show overestimation and underestimation of the pouring angle compared with the ground truth. These two error patterns may indicate different strategies of physical reasoning. To examine whether there is a switching mechanism between the two patterns among those conditions, we employed symbolic regression to automatically identify an explainable factor and its corresponding switching point that best distinguishes between the two patterns. We considered all experimental design factors, including cup shape, object shape, object size, and filling height, along with the object number and simulated pouring angle. Our analysis shows that the simulated pouring angle effectively differentiates between the reversal patterns observed in human participants' estimations of tilt angles for pouring (see Fig. 1C). We identified the optimal boundary for distinguishing these patterns to be 65 degrees by searching from 20 to 120 with an interval of 1. Initially, participants tended to overestimate these angles when the simulated pouring angles were relatively small (mean discrepancy = 7.76 ± 13.67). As the angle increases, this trend shifts to consistent underestimation (mean discrepancy = -9.89 ± 8.75). Given the consistent tilting speed, the observed pattern switch as the pouring angle increases suggests a hypothesis that the physical reasoning strategy may change when the simulation time exceeds a certain resource boundary.

IPE fails to explain all trials

To validate our hypothesis, we first experiment with the IPE model. Fitting human judgments in the overestimation phase with IPE supports our hypothesis of the simulation strategy's dominance in the shorter time span. Note that as the angular speed remains constant in our experiments, the simulation time is proportional to the degree of angle. When the positional noise and rotational speeds of the IPE model were optimized, the results were closely aligned with human performance, explaining the overestimation pattern effectively (r = .890).

However, once the pouring angle exceeded the 65-degree boundary, IPE's prediction error significantly increased (t(52) = -3.354, p = .002; see Fig. 2b for absolute error comparison on the left). No parameter combination in the IPE model could well explain the underestimation pattern, indicating the existence of an alternative strategy other than IPE.

Table 1: **Categories, coefficients, and p-values of physical variables in the learned heuristic model.** All physical variables except the object shape show significant contributions to the outcomes.

Variable	Category	Coefficients	p
Cup shape	H-shape	-11.528	0.000
	A-shape		
	V-shape		
Object shape	Circle	1.577	0.073
	Triangle		
	Trapezoid		
Object size	Small	7.029	0.000
	Medium		
	Large		
Filling height	Half	-19.955	0.000
	Full		

Learned heuristic model complements IPE

To better explain the underestimation pattern in human behavior, we devised a heuristic model incorporating key physical attributes rooted in our experiments: filling height, cup shape, object shape, and object size. This model effectively compensated for the discrepancies unexplained by the IPE model. The heuristic model performed well when the actual pouring angle exceeded 65 degrees (r = .841), but its accuracy diminished below this boundary (Mann-Whitney U test, p = .003; see Fig. 2b for absolute error comparison on the right).

Further analysis of specific heuristics revealed that filling height, cup shape, and object size significantly influence heuristic judgment (see Tab. 1, p = .000 for all three variables). The model's coefficients allowed a quantitative assessment of these variables' impact. Despite the simplicity and approximate encoding, this linear heuristic model captured basic physical intuition effectively. The findings align with our second hypothesis, suggesting the adoption of heuristic strategies when mental simulation reaches its boundary.

SHM explains human judgments on all conditions

Building upon our findings, we constructed the Simulation-Heuristics Model (SHM), a dual-process model integrating both simulation and heuristic strategies, to optimally predict human performance across all trials. Instead of relying on actual simulation time in humans, which is unavailable, we instead based the transition criterion in SHM on IPE's simulation time. A grid search identified the boundary of 68.2 degrees in simulation time and a dynamic positional noise of 0.2 as optimal for mirroring human judgments.

In predicting overall human performance, SHM surpassed three baseline models: the deterministic physics model, IPE, and the purely heuristic model. SHM exhibited the highest correlation and lowest RMSE (r = .834, RMSE = 10.002), as shown in Fig. 3. Although IPE was correlated with human judgments (r = .772), it showed high error in making human-like predictions (RMSE = 17.457). On the contrary, the heuris-
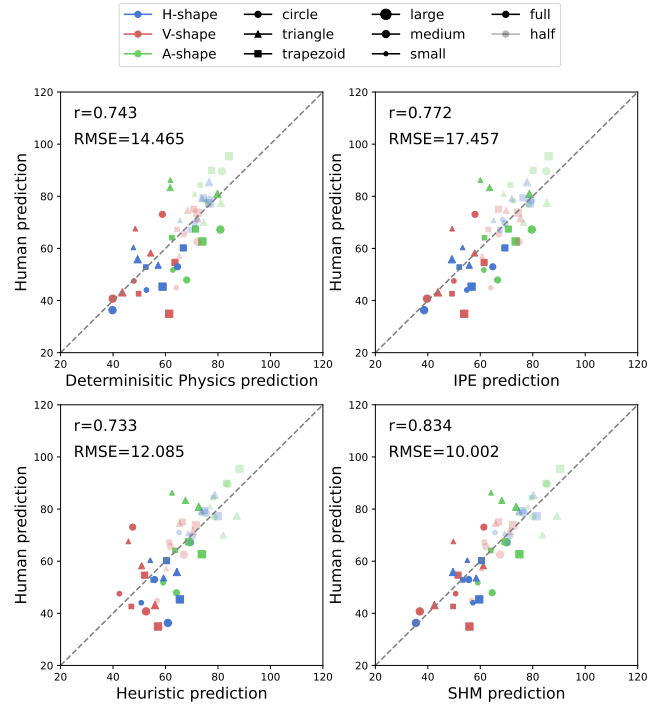

Figure 3: **Comparison between SHM and other baseline models.** The correlation and RMSE between model predictions and human predictions across all 54 conditions are compared. Among the four models evaluated, SHM demonstrates the highest correlation and the lowest RMSE, indicating its superior predictive accuracy.

tic model could predict human judgments with smaller RMSE but failed to better explain the variance (r = .733, RMSE = 12.085).

The fitted SHM model exhibited consistent predictive performance across diverse conditions (*e.g.*, different cup shapes, object shapes, sizes, and filling heights). It consistently showed the lowest RMSE, except in specific scenarios where the heuristic model was parallel (Fig. 4). The model explained maximum variance in almost all cases, with comparable performance to IPE in scenarios involving large or trapezoidal marbles. Notably, in situations where IPE minimally correlated with human judgments (*e.g.*, A-shaped cups, r = .461), SHM maintained effectiveness (A-shaped cups, r = .647). It also significantly improved correlation in scenarios poorly addressed by the heuristic model (full filling height, r improved to .673 from .377). These results highlight SHM's capability to synergize the strengths of both IPE and the heuristic model, enabling robust predictions across diverse scenarios. Consequently, the SHM model, with its transition mechanism based on simulation time, aligns with our third hypothesis and effectively accounts for a wide range of conditions and metrics.

Discussion

Limitations While our study proposes a method to distinguish between cognitive strategies, people can employ multiple strategies within a single trial. Furthermore, existing
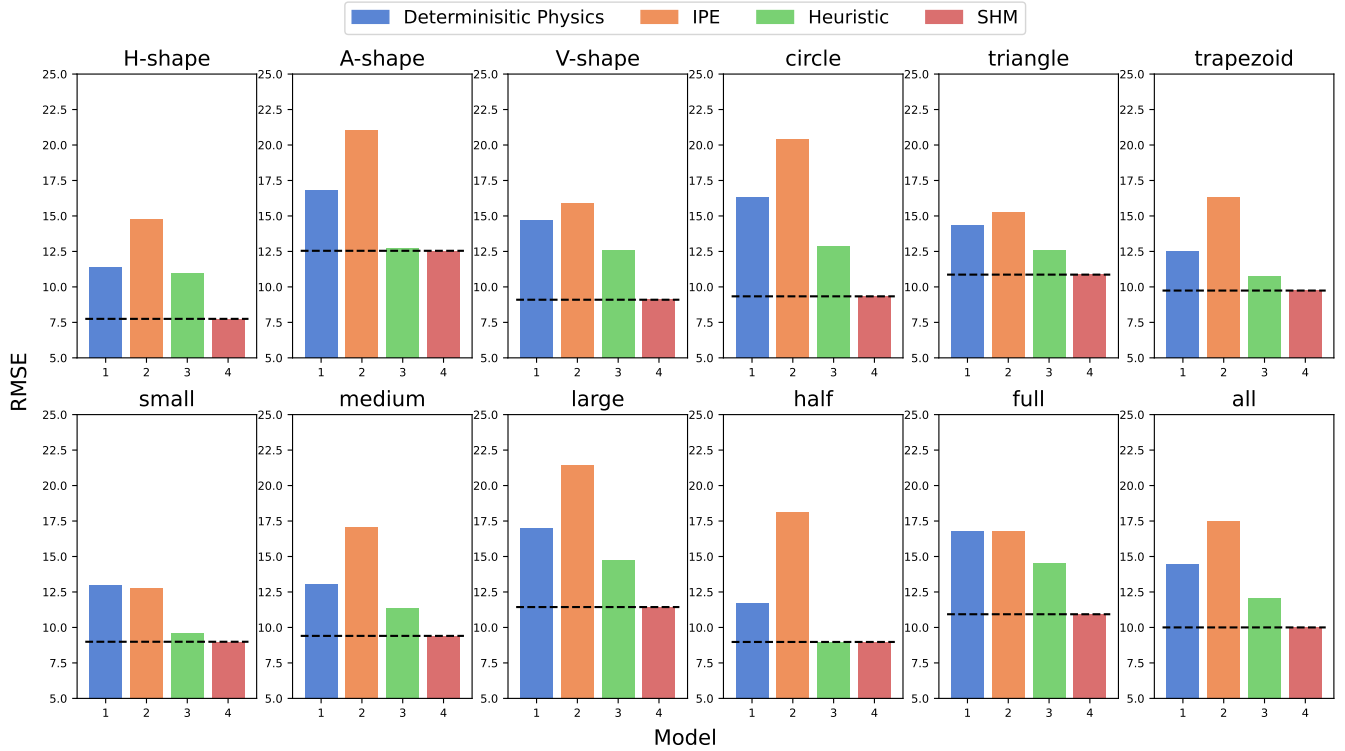
Figure 4: **Comparison of four models' RMSE on different conditions.** RMSE is calculated as the root mean square error between the model's predicted pouring angle and the human judgments. The bottom right figure represents the performance across all 54 trials. A dashed line is included to indicate the RMSE of the SHM, showing a clear advantage when compared with other models.

research (Siegler, 1976; Marcus & Davis, 2013; Davis & Marcus, 2016; Ludwin-Peery et al., 2021) in intuitive physics indicates that humans' results can be incompatible with those from mental simulation models, even for simple tasks like predicting the tipping direction of the balanced beams. This suggests that factors other than the mental cost triggered the heuristic strategy, and a more nuanced modeling is needed to capture the variability in human judgment. We note that our model parameters were fitted using the complete dataset rather than through cross-validation on separate subsets of conditions. Therefore, our claims focus on the model's ability to explain human behavior consistently across diverse physical scenarios rather than its ability to predict novel conditions.

Future work Future research could investigate the mechanisms behind strategy selection and the generalizability of our findings across various physical scenarios. Key questions include how humans choose an initial strategy and estimate simulation costs. Eye-tracking or verbal protocol studies might reveal whether people start with simulation and abandon it when costly, or make preemptive decisions based on scene complexity. Although our use of pouring angles is task-specific, the underlying cost-based strategy-switching mechanism likely applies more broadly.

Additionally, the dual-process model employed by humans may utilize different physical variables as heuristics across scenarios. For instance, tower height might serve as a heuristic in block collapse estimation, while object distribution could be

relevant for predicting group motion. Other factors influencing strategy selection might include task complexity, scenario familiarity, and prior experience. Despite this variance in heuristic strategies, investigating whether our learning approach maintains effectiveness across contexts would be valuable for developing a deeper understanding of physical reasoning.

Conclusion

In this work, we design a pouring-marble task to study the computational mechanism in intuitive physics. The sequential experiments underscore that while the IPE effectively predicts human judgments in scenarios with short simulation times, its efficacy diminishes as these times extend. This limitation of IPE paves the way for the implementation of a heuristic approach that shows greater accuracy in scenarios necessitating longer and more complex simulations. The introduction of the SHM model, which seamlessly integrates these two cognitive strategies based on the simulation cost, not only aligns more closely with human behavior but also enhances the model's consistent predictive capabilities across varied conditions. By bridging the gap between mental simulations and heuristic approaches, the SHM model offers a robust framework to capture the complexity and adaptability of human cognition in intuitive physics. This model serves as a pivotal step in exploring computational methods that closely mimic human-like reasoning, providing insights into the cognitive mechanisms that govern our interactions with the physical world.

References

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

Baillargeon, R., & Graber, M. (1987). Where's the rabbit? 5.5-month-old infants' representation of the height of a hidden object. *Cognitive Development*, *2*(4), 375–392.

Baillargeon, R., Spelke, E. S., & Wasserman, S. (1985). Object permanence in five-month-old infants. *Cognition*, *20*(3), 191–208.

Balaban, H., & Ullman, T. (2024). *The capacity limits of mental simulation.* PsyArXiv.

Bass, I., Smith, K. A., Bonawitz, E., & Ullman, T. D. (2021). Partial mental simulation explains fallacies in physical reasoning. *Cognitive Neuropsychology*, *38*(7-8), 413–424.

Bates, C., Battaglia, P. W., Yildirim, I., & Tenenbaum, J. B. (2015). Humans predict liquid dynamics using probabilistic simulation. In *Annual meeting of the cognitive science society (cogsci)*.

Bates, C. J., Yildirim, I., Tenenbaum, J. B., & Battaglia, P. (2019). Modeling human intuitions about liquid flow with particle-based simulation. *PLoS computational biology*, *15*(7), e1007210.

Battaglia, P., Ullman, T., Tenenbaum, J., Sanborn, A., Forbus, K., Gerstenberg, T., & Lagnado, D. (2012). Computational models of intuitive physics. In *Annual meeting of the cognitive science society (cogsci)*.

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 (PNAS)*, *110*(45), 18327–18332.

Chen, T., Allen, K. R., Cheyette, S. J., Tenenbaum, J., & Smith, K. A. (2023). " just in time" representations for mental simulation in intuitive physics. In *Annual meeting of the cognitive science society (cogsci)*.

Davis, E. (1998). Naive physics perplex. *AI magazine*, *19*(4), 51–51.

Davis, E., & Marcus, G. (2016). The scope and limits of simulation in automated reasoning. *Artificial Intelligence*, *233*, 60–72.

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

Gerstenberg, T., Zhou, L., Smith, K. A., & Tenenbaum, J. B. (2017). Faulty towers: A hypothetical simulation model of physical support. In *Annual meeting of the cognitive science society (cogsci)*.

Guevara, T. L., Taylor, N. K., Gutmann, M., Ramamoorthy, S., & Subr, K. (2017). Adaptable pouring: Teaching robots not to spill using fast but approximate fluid simulation. In *1st conference on robot learning 2017*.

Kahneman, D. (2011). *Thinking, fast and slow*. macmillan.

Kaiser, M. K., Jonides, J., & Alexander, J. (1986). Intuitive reasoning about abstract and familiar physics problems. *Memory & Cognition*, *14*, 308–312.

Kim, I. K., & Spelke, E. S. (1992). Infants' sensitivity to effects of gravity on visible object motion. *Journal of Experimental Psychology: Human Perception and Performance*, *18*(2), 385.

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

Kubricht, J., Jiang, C., Zhu, Y., Zhu, S.-C., Terzopoulos, D., & Lu, H. (2016). Probabilistic simulation predicts human performance on viscous fluid-pouring task. In *Annual meeting of the cognitive science society (cogsci)*.

Kubricht, J., Zhu, Y., Jiang, C., Terzopoulos, D., Zhu, S., & Lu, H. (2017). Consistent probabilistic simulation underlying human judgment in substance dynamics. In *Annual meeting of the cognitive science society (cogsci)*.

Kubricht, J. R., Holyoak, K. J., & Lu, H. (2017). Intuitive physics: Current research and controversies. *Trends in Cognitive Science*, *21*(10), 749–759.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, *40*, e253.

Li, Y., Wang, Y., Boger, T., Smith, K. A., Gershman, S. J., & Ullman, T. D. (2023). An approximate representation of objects underlies physical reasoning. *Journal of Experimental Psychology: General*.

Lopez-Guevara, T., Pucci, R., Taylor, N. K., Gutmann, M. U., Ramamoorthy, S., & Suhr, K. (2020). Stir to pour: Efficient calibration of liquid properties for pouring actions. In *International conference on intelligent robots and systems (iros)*.

Ludwin-Peery, E., Bramley, N. R., Davis, E., & Gureckis, T. M. (2020). Broken physics: A conjunction-fallacy effect in intuitive physical reasoning. *Psychological Science*, *31*(12), 1602–1611.

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

Marcus, G. F., & Davis, E. (2013). How robust are probabilistic models of higher-level cognition? *Psychological science*, *24*(12), 2351–2360.

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.

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.

Piloto, L. S., Weinstein, A., Battaglia, P., & Botvinick, M. (2022). Intuitive physics learning in a deep-learning model inspired by developmental psychology. *Nature Human Behaviour*, *6*(9), 1257–1267.

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

Shepard, R. N., & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science*, *171*(3972), 701–703.

Siegler, R. S. (1976). Three aspects of cognitive development. *Cognitive psychology*, *8*(4), 481–520.

Smith, K., Hamrick, J., Sanborn, A. N., Battaglia, P., Gerstenberg, T., Ullman, T., & Tenenbaum, J. (2024). *Intuitive physics as probabilistic inference*. MIT Press.

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., Peres, F., Vul, E., & Tenebaum, J. (2017). Thinking inside the box: Motion prediction in contained spaces uses simulation. In *Annual meeting of the cognitive science society (cogsci)*.

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

Ullman, T., & Wang, Y. (2023). Resource bounds on mental simulations: Evidence from a fluid-reasoning task.

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.

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

Xu, K., Srivastava, A., Gutfreund, D., Sosa, F., Ullman, T., Tenen-
    baum, J., & Sutton, C. (2021). A bayesian-symbolic approach to
    reasoning and learning in intuitive physics. In *Advances in neural
    information processing systems (neurips)*.

Yildirim, I., Gerstenberg, T., Saeed, B., Toussaint, M., & Tenen-
    baum, J. (2017). Physical problem solving: Joint planning with
    symbolic, geometric, and dynamic constraints. *arXiv preprint
    arXiv:1707.08212*.

Zhou, L., Smith, K. A., Tenenbaum, J. B., & Gerstenberg, T. (2023).
    Mental jenga: A counterfactual simulation model of causal judg-
    ments about physical support. *Journal of Experimental Psychol-
    ogy: General*, *152*(8), 2237.