MOTIVATIONAL INFLUENCES ON INTUITIVE PHYSICAL JUDGMENTS

by

Meriel Doyle

Dr. Yuan Chang Leong, Advisor

Dr. Akram Bakkour, Reader

A Thesis Submitted in Partial Fulfillment of the Requirements for the

Degree of Bachelor of Arts with Honors in Psychology

UNIVERSITY OF CHICAGO

May 2023

ABSTRACT

People rely extensively on their inferences about physical scenarios when interacting with the world around them, but examples of flawed physical inference are well-documented (e.g., real-world optical illusions). In the current work, we test the hypothesis that our intuitive physical judgments can change when we are motivated to see a particular outcome. In a Prolific study (N = 70) and an in-lab eye-tracking study (N = 28), participants were shown images of 3D block towers and rewarded for correctly judging whether each tower would fall or remain standing under the influence of gravity. We motivated participants by offering bonuses if towers had particular stabilities—though crucially, bonuses depended on objective tower stability, not participants' responses, making accuracy the optimal strategy. Despite this, we found that, on average, participants were more likely to judge towers as stable when motivated to see them as stable, and more likely to judge towers as unstable when motivated to see them as unstable. Specifically, motivation appeared to bias participants' physical judgments against an intrinsic response bias (e.g., a general tendency to indicate towers as stable or unstable). Moreover, eye-tracking revealed that fixation patterns were more similar between participants making the same judgement, as well as participants motivated to see the same outcome. These results suggest that *how* people sample information may impact their physical judgments, and that motivation may *influence* how people sample this information in order to make these judgments. Future work will compare human physical inference strategies with simulation-based physics engine models and Inception v4 convolutional neural network (CNN) architectures. The results of this project may shed light on the potential limitations of human physical scene understanding and, ultimately, inform the development of more human-like artificial intelligence systems.

INTRODUCTION

To what extent can we trust our instincts about the world around us? A majority of our daily activities rely on physical intuition: from simple tasks like brushing our teeth or tying our shoelaces, to more complex ones like driving a car. While we are generally able to make quick judgments about the physical relationships between different objects in our environment, this task becomes harder as scenes grow more complex. For example, accurately assessing the exact position of each wooden block in a Jenga tower relative to its neighbors can be challenging, and although we can make estimations, incorrect evaluations of the tower's stability may lead to its collapse. Indeed, it is well-known that humans are prone to making false inferences about objects in their visual field (e.g., optical illusions). Understanding how motivations, desires, prior knowledge, and/or belief systems shape perception is therefore crucial. The current study investigates whether external motivational factors, such as rewards, risks, and uncertainty, impact the way we perceive, reason, and make decisions about our physical environment.

Motivated perception, or "motivated seeing," is an area of neuroscience research that examines the influence of motivation on perception and decision-making. The phrase "motivated seeing" refers to the idea that individuals perceive the world around them based on desires, goals, or rewards, in turn, leading to biased interpretations of sensory information. In a recent study on motivated perception, Leong et al. (2019) used financial incentives to influence people's perceptual judgments. Integrating results from neuroimaging approaches (e.g., fMRI) and computational modeling (e.g., drift diffusion models), they found that motivation to see a certain percept over another during a visual categorization task increased neural activity specific to a motivationally relevant visual category. This bias in participants' *neural representation* of

the presented image provides a computational explanation for how reward leads to inaccurate world representations—participants were literally "seeing" what they wanted to see. Our study extends this finding to examine how motivational influences might impact the way people intuitively reason about, represent, and interact with their *physical* environment.

Intuitive physical inference is central to the human experience and a defining feature of human intelligence, yet we lack a comprehensive account of the supporting mental processes and computations. Cognitive science suggests that physical intuition depends on the formation of "mental models"—representations of the objects in a given scenario and their relationships to each other, according to the rules of physics (Johnson-Laird, 2010). The leading framework proposes that these models operate through "mental simulation," a flexible cognitive process that involves the internal reconstruction or imagination of a scenario, enabling people to interpret complex situations, problem-solve, and adapt to their ever-changing environment (Davis & Marcus, 2015). In the context of physical reasoning, it is hypothesized that people mentally simulate the behavior of objects in a system in order to make judgments and predictions about their stability, motion, or interactions. However, our limited understanding of these processes is evidenced by our inability to replicate any human-like computational adaptability in our modern artificial intelligence (AI) systems (Ludwin-Peery et al, 2021).

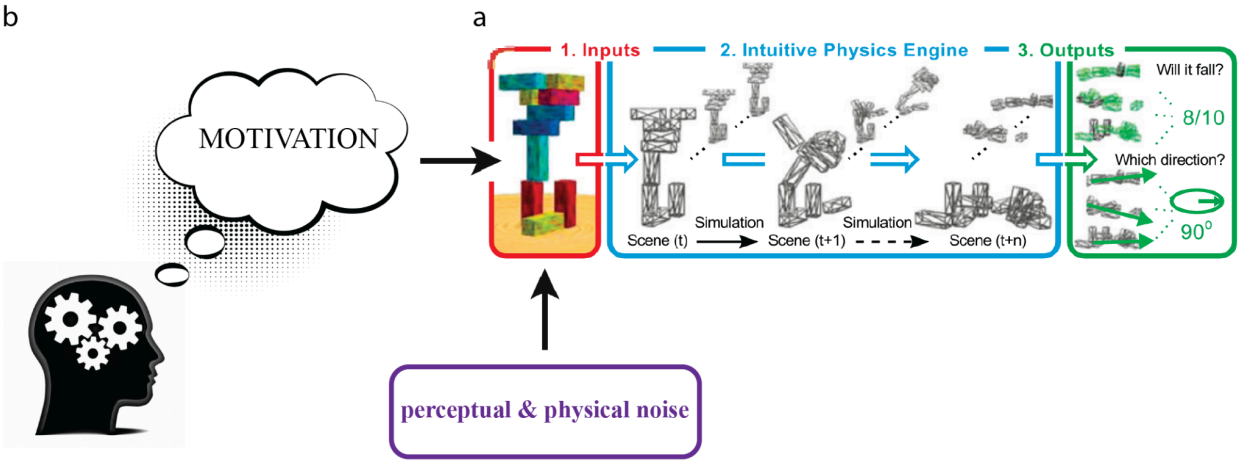

Fig. 1 | The intuitive physics engine (IPE). a, Adapted from Battaglia et al. (2013): It is hypothesized that people reason about physical scenes in the following way: (1) As inputs, the model takes an observer's internal reconstruction of a scene, modulated by a set of *noise* parameters (e.g. position, velocity, mass, friction, etc.) to account for the perceptual and physical uncertainties that an observer has about the scene. (2) The model then applies the relevant laws of Newtownian physics over a distribution of these inputs, to simulate how each scene in the distribution will unfold over time. (3) Finally, the model outputs the scenario's most likely outcome. **b**, In theory, the IPE architecture interfaces flexibly with lower-level perceptuomotor systems *and* higher-level cognitive systems responsible for planning, action, reasoning, and language. Therefore, model outputs (i.e., human physical judgments) may be influenced, at least in part, by these higher-level cognitive systems. For example, an observer's prior knowledge or belief systems might impact the degree of uncertainty their IPE incorporates. Relatedly, it may be possible for *motivational* influences such as risk and reward to influence the salience of certain features an observer focuses on when reconstructing a mental model of the scene. The present study explores the ways in which motivational influences, specifically, impact physical scene understanding.

Ullman et al. (2017) proposed an account of mental simulation that can be instantiated using computer models. These models represent human physical judgments with an "Intuitive Physics Engine" (IPE), a system analogous to video game physics engines (Fig. 1*a*). The IPE takes an observer's internal reconstruction of a scene (i.e., their "mental model"), modulated by a set of "uncertainty" noise parameters, and runs simulations applying Newtonian physics across the distribution of inputs. It outputs statistically most likely outcomes—for example, when a person wants to decide whether a 3D block tower will fall or remain standing, their IPE would output the average proportion of blocks that fell across the set of simulation results.

The predictions generated by these physics engine models correlate strongly with human judgments across an array of tasks, such as assessing block tower stability and predicting fall direction (Battaglia et al., 2013). However, further research is needed to more precisely define the interplay between visual features, mental simulations, and other heuristic principles in human physical inference (Zhou et al., 2021). For example, while Battaglia et al. (2013)'s IPE assumes that people run noisy, approximate simulations, Hamrick et al. (2015) provide evidence that people vary simulation frequency based on uncertainty, following the sequential probability ratio test (SPRT; also known as the drift-diffusion model). Decision strategies thus may depend on specific task goals and characteristics. Whether certain features constrain the use of simulation, and what alternatives the mind employs, remain open questions (Ludwin-Peery et al., 2021).

Crucially, the IPE architecture predicts flexible integration with perceptuomotor and higher-level cognitive systems (Fig. 1*b*). Therefore, in the context of the current study, we might expect motivation to directly affect intuitive physical inference at the level of the IPE inputs (i.e., the observer's internal reconstruction of a scene). We can explore this effect using eye-tracking fixation and saccade pattern data. For example, when motivated toward particular outcomes, do people preferentially sample certain visual information? Do people with the same goals sample more similar information than people with opposing goals? Relatedly, how might these modulations at the input level impact the occurrence and/or frequency of mental simulation? Do model outputs (i.e., the observer's predicted outcome of the scenario) support approximate, probabilistic versus SPRT accounts of mental simulation?

The goal of the present study was thus twofold: (1) to extend the "motivated perception" literature to the domain of physical inference by testing whether financial incentives bias physical judgments (Study 1), and (2) to use eye-tracking to measure how motivation affects information sampling (Study 2). Both studies used a visual categorization task motivating participants to perceive 3D block towers as more/less stable. For Study 1, we predicted that participants would be more likely to categorize an ambiguous image as the category they were motivated to see, and for Study 2, we predicted that for a given block tower, fixation patterns would be more similar between participants who predicted the *same* outcomes, as compared to participants who predicted opposing outcomes.

Additionally, we aimed to compare human physical reasoning strategies with those of AI systems more generally, in order to improve our understanding of the underlying mechanisms in both. Our eye-tracking analysis therefore compared human strategies with those of a deep learning visual classifier trained on the ShapeStacks dataset—20,000 towers with objective Fall/Stand rankings derived from training a convolutional neural network (CNN) architecture, Inception v4, to distinguish between stable/unstable structures (Groth et al., 2018) (Fig. 2).

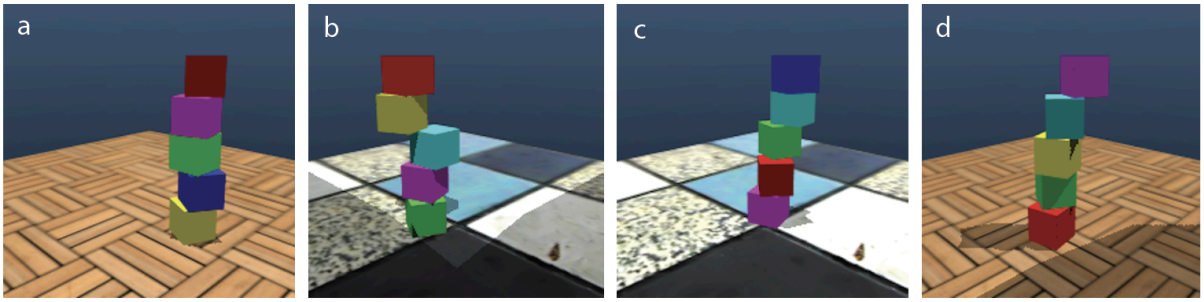

Fig. 2 | Different scenarios from the ShapeStacks dataset. (**a** - **d**) depict initial stack setups: **a**, Stable, rectified tower of cubes. **b**, Unstable tower where multiple objects counterbalance each other. **c**, Stable, but visually challenging scenario due to colors and textures. **d**, Unstable, visually challenging scenario due to colors and textures. Towers depicted in **b** and **d** collapse due to a center of mass violation (VCOM). While the ShapeStacks dataset contains towers with many kinds of structural stability violations, only towers with VCOM instabilities were used in our study.

The towers include annotations of mechanical failure points—blocks whose removal would cause the tower to collapse (Fig. 3*c*). To determine whether the network depended on these points to make its judgments, Groth et al. (2018) conducted an "occlusion study," where they partially obscured image regions to test network dependence (Fig. 3*a*). If occlusion changed the network's stability prediction, it was inferred that the network relied on that part of the image to make its prediction. Using this approach, the authors generated "attention maps" for each image, highlighting the parts that contributed to the increase or decrease in stability ratings for the towers—in other words, the parts that the network had *learned* to attend to (Fig. 3*b*). Network "attention" was strongly correlated with mechanical failure points, indicating that the network had learned to attend to structural instabilities when making its judgments (Fig 3*b*, *c*).

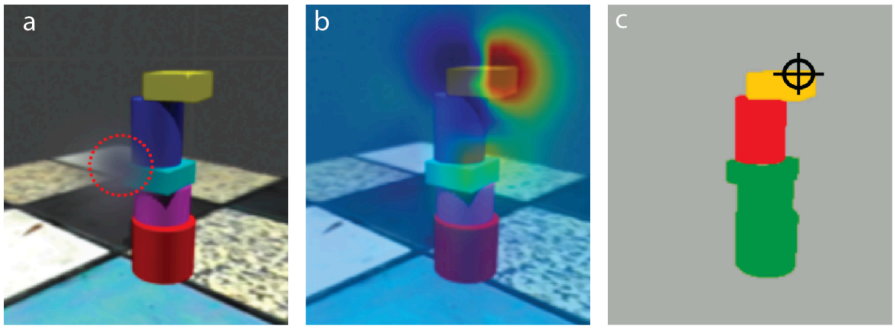

Fig. 3 | ShapeStacks dataset: Inception v4 CNN attention heatmap obtained via occlusion study (Groth et al. 2018). a, A blur is applied to the image. **b**, The increase (red) / decrease (blue) in the predicted stability is shown as a heatmap. **c**, The heatmap is compared to the ground-truth segmentation maps. **c**, violation sites as indicated by the crosshairs are correlated with the centers of the attention

Given the CNN's focus on stability violations, we tested whether humans, too, might intuitively employ similar strategies. Using fixation data, we examined whether people attend to the same failure points, as well as the motivational context in which they do so. In general, we predicted that participants motivated to see stability would attend *less* to *instability* points, while those motivated to see instability would focus *more* on these points (perhaps in order to mentally simulate the ways in which the tower might fall).

STUDY 1 – MOTIVATION

Methods

Participants. 71 participants were recruited from the Prolific platform (www.prolific.co) [2022]. Prolific recruitment was restricted to participants in the US who were fluent in English and had normal or corrected-to-normal vision. Recruitment was also restricted to participants with an approval rating of at least 98% and between 100 and 10,000 previous submissions. Data from 1 participant was discarded due to a lack of visible effort in the task, yielding an effective sample size of 70 participants (23 male, 47 female, 19-70 years of age, mean age = 37.7 years). All experimental procedures were approved by the University of Chicago Institutional Review Board. Participants were provided written, informed consent before the start of the study and were paid between $12 and $20 depending on their task performance.

Stimuli. For each experimental block, participants were presented with the same selection of 40 ShapeStacks tower images, which varied both in terms of objective stability and judgment difficulty: 10 unambiguously stable (e.g., “easy stand”), 10 unambiguously unstable (e.g., “easy fall”), 10 ambiguously stable (e.g., “hard stand”), and 10 ambiguously unstable (e.g., “hard fall”). The experiment was coded in JavaScript using JsPsych (de Leeuw, 2015) and hosted on Cognition (www.cognition.run) [2022].

Behavioral Task. Participants completed three 40-trial blocks (neutral, cooperation, competition). In neutral condition trials, participants were presented with a tower image and given 5 seconds to predict either ‘FALL’ or ‘STAND’ by responding with the ‘right’ arrow (►) or ‘left’ arrow (◄) on their keyboard. In cooperation/competition conditions (Fig. 4*a*), a ‘teammate’ or ‘opponent’ bet ‘FALL’ or ‘STAND’ on each image before participants responded. Correct teammate bets earned both parties bonuses; incorrect lost money. Competition reversed this: correct opponent bets cost participants money; incorrect earned bonuses. Crucially, the reward-maximizing strategy was ignoring bets and prioritizing accuracy (Fig. 4*b*). Including conditions where participants responded both consistently and inconsistently with the presented bet controlled for semantic priming—ensuring responses reflected motivated perception rather than priming from having just seen the word ‘FALL’ or ‘STAND’.

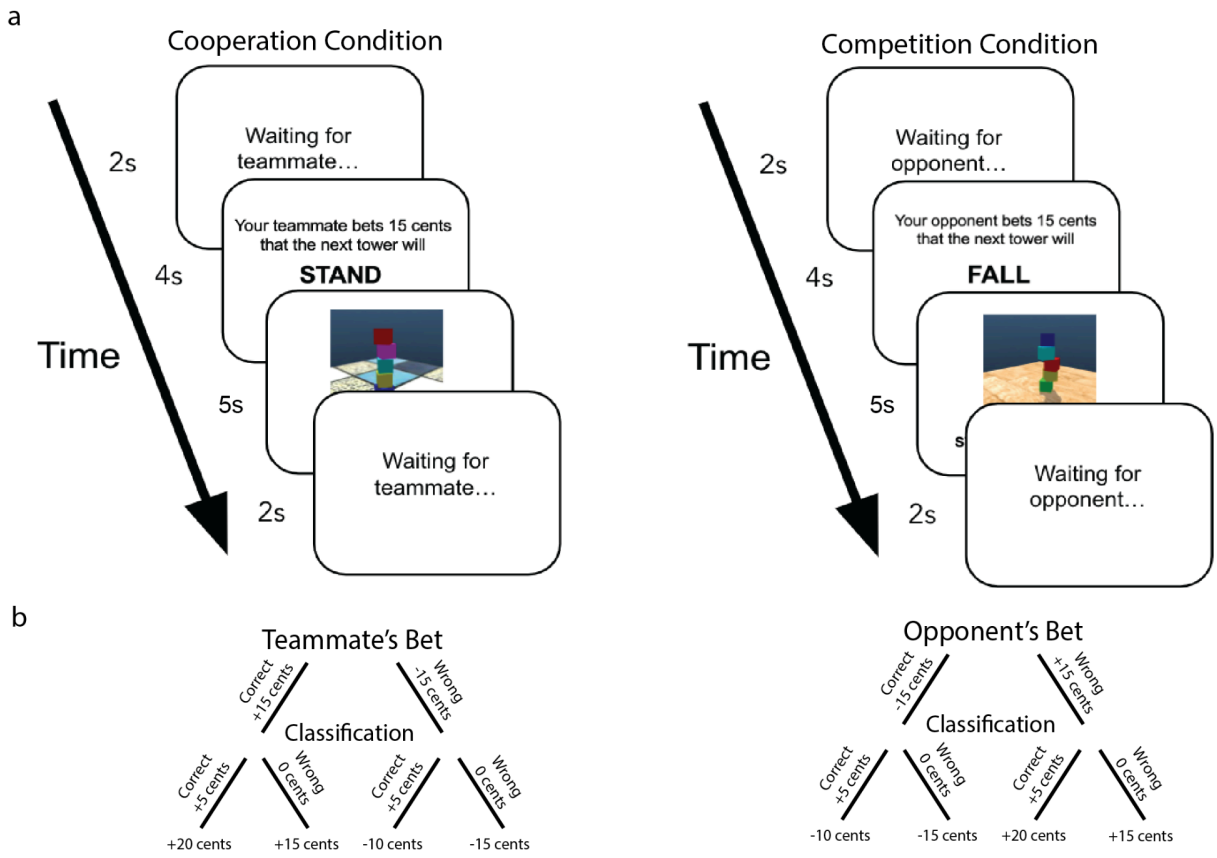

Fig. 4 | Experimental design. a, Participants were presented with ShapeStacks tower images. In the cooperation condition, a teammate first makes a bet about whether the tower will fall or stand. Participants are then presented the image and have to similarly determine whether the tower will fall or stand. In the competition condition, an opponent makes the bet instead. **b**, Payoff structure. Participants won 5 cents for each correct categorization. They won an extra 15 cents if the teammate's bet was correct, but lost 15 cents if the teammate's bet was wrong. Conversely, participants lost 15 cents if the oppoent's bet was correct, but won 15 cents if the opponent's bet was wrong. As the outcome of the bets was determined by the objective Fall/Stand ranking of the presented image, and not by participants' subjective categorizations, the reward maximizing strategy was to ignore the bets and perform the categorizations accurately.

Psychometric functions. We employed Generalized Linear Mixed Effects (GLME) models to estimate motivation's impact on categorizations. One model included all conditions (i.e., cooperation, competition, neutral) and participants. Other models were fit to data from all conditions, but for two particular participant subgroups—one with an inherent bias to respond 'FALL,' and the other with an inherent bias to respond 'STAND'. Models incorporated tower stability (% stand) as a covariate, with random intercepts and slopes for motivation effects. Analyses used glmer (lme4 package) with Satterthwaite approximation (Bates et al., 2015; Kuznetsova et al., 2017). A *logit* link function was used. Models were specified as follows, with random effects indicated in parentheses:

$$response \sim \%\, stand + motivation + (motivation \mid subj) \qquad (1)$$

Results

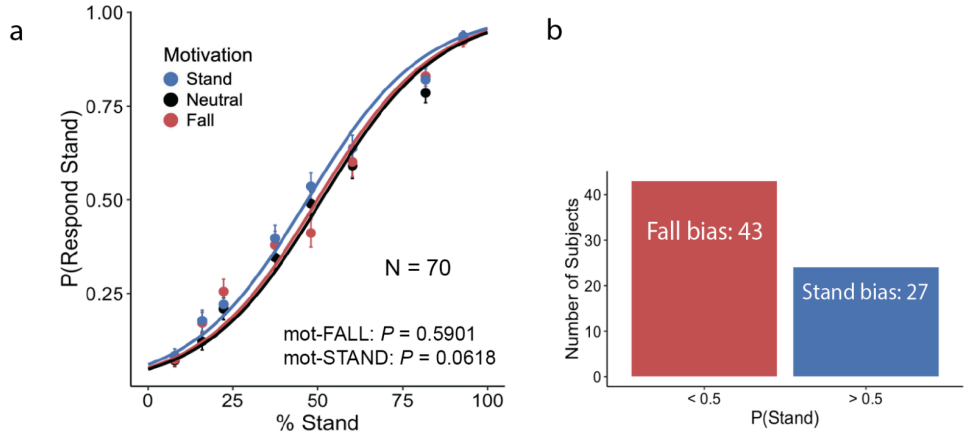

Fig. 5 | Motivation biases intuitive physical judgements. Pannels **a** and **b** contain data from 70 participants. **a**, Participants were more likely to report seeing the category they were motivated to see. When participants were motivated to see the tower stand (e.g. mot-STAND, the teammate bet 'STAND' or the opponent bet 'FALL'), their psychometric function was shifted left. The inverse pattern was not obvious for situations where participants were motivated to see the tower fall (e.g. mot-FALL, the teammate bet 'FALL' or the opponent bet 'STAND'). Statistical significance for these psychometric functions, as well as those depicted in Fig. 6, 8, & 9, was assessed using GLME (see 'Psychometric functions' in Methods). Error bars indicate s.e.m. **b**, 43 participants had a bias for responding 'FALL', and 27 had a bias for responding 'STAND'.

Motivation biases intuitive physical judgments. As expected, in the neutral condition, participants accurately tracked tower stability, responding 'STAND' more frequently as towers became objectively more stable (GLME: $z = 45.5$, $p < .001$, $\boldsymbol{b = 0.059}$, SE = 0.0013).We then estimated separate psychometric functions for participants motivated to see the tower stand (i.e., either the teammate bet 'STAND' or the opponent bet 'FALL') and participants motivated to see the tower fall (i.e., either the teammate bet 'FALL' or the opponent bet 'STAND') (Fig. 5*a*). On average, we found that, compared to the proportion of participants who responded 'STAND' in the neutral condition, participants were (insignificantly) more likely to respond 'STAND' when they were motivated to see the tower stand (GLME: $z = 1.87$, $p = .0618$, $\boldsymbol{b = 0.25}$, SE = 0.13) and (insignificantly) more likely to respond 'STAND' when they were motivated to see the tower fall (GLME: $z = 0.54$, $p = .5901$, $\boldsymbol{b = 0.074}$, SE = 0.14).

This unexpected pattern—that participants responded 'STAND' more frequently when motivated to see the tower stand *and* fall—prompted further analysis. Interestingly, we found that individual participants tended to exhibit a bias for indicating towers as stable or unstable. Specifically, 43/70 participants exhibited a 'STAND' response bias (>50% 'STAND' responses), and 27/70 exhibited a 'FALL' response bias (>50% 'FALL' responses) in the neutral condition (Fig. 5*b*). We therefore decided to examine the effect of motivation for each of these two participant subgroups separately.

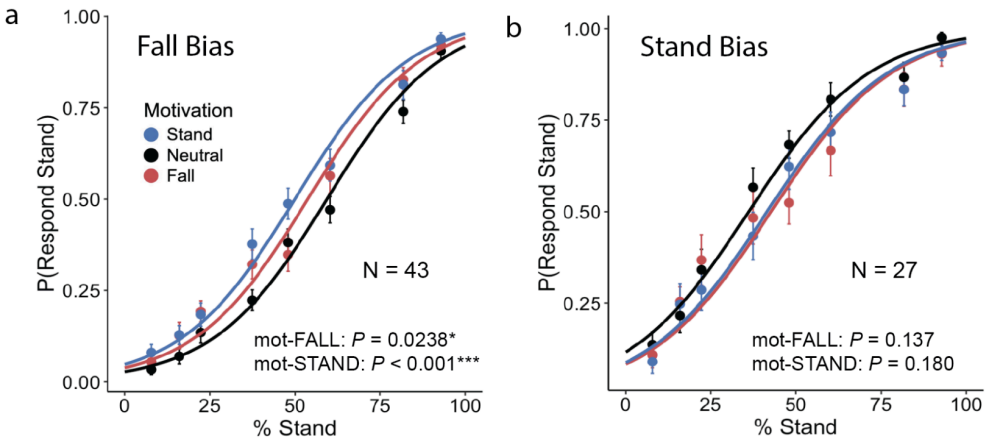

Fig. 6 | Motivation biases physical judgements against an intuitive response bias. a, Fall-bias participants were more significantly likely to be influenced by motivation to see the tower **stand** (e.g. mot-STAND) than motivation to see the tower fall. **b**, Stand-bias participants were (insignificantly) more likely to be influenced by motivation to see the tower **fall** (e.g. mot-FALL) than motivation to see the tower stand.

Motivation biases physical judgments against an intuitive response bias. Both subgroups accurately tracked tower stability in neutral conditions (fall-bias GLME: $z = 37.09$, $p < .001$, $\boldsymbol{b = 0.060}$, SE = 0.0016; stand-bias GLME: $z = 26.1$, $p < .001$, $\boldsymbol{b = 0.055}$, SE = 0.0021). When calculating separate psychometric functions for each subgroup, we found that fall-bias participants were significantly more likely to respond 'STAND' when they were motivated to see the tower stand (GLME: $z = 4.05$, $p < .001$, $\boldsymbol{b = 0.56}$, SE = 0.14) and (insignificantly) more likely to respond 'STAND' when they were motivated to see the tower fall (GLME: $z = 2.26$, $p = .024$, $\boldsymbol{b = 0.34}$, SE = 0.15) (Fig 6*a*). Conversely, stand-bias participants were (insignificantly) less likely to respond 'STAND' when they were motivated to see the tower stand (GLME: $z = -1.34$, $p = .180$, $\boldsymbol{b = -0.30}$, SE = 0.23) and (insignificantly) less likely to respond 'STAND' when they were motivated to see the tower fall (GLME: $z = -1.37$, $p = .137$, $\boldsymbol{b = -0.37}$, SE = 0.25) (Fig. 6*b*).

Taken together, these results provide support for our initial hypothesis that participants would be more likely to categorize an ambiguous image as the category they were motivated to see. The GLME models for the participant subgroups, however, indicate that a more nuanced interpretation is warranted. Indeed, motivation appears to have encouraged participants to respond *in opposition to their inherent response biases*. Moreover, this effect was stronger for fall-bias participants—in general, their judgments were more likely to be influenced by motivation to see the tower *stand* than motivation to see the tower fall. While a similar pattern did emerge for stand-bias participants, they were only slightly (and insignificantly) more likely to be influenced by motivation to see the tower *fall* than motivation to see the tower stand. These results were both interesting and unexpected, so as we moved on to a similar, in-lab version of the study with an eye-tracking component, we kept this finding in mind.

STUDY 2 – EYE-TRACKING PHYSICAL INFERENCE

Methods

Participants. 60 participants were recruited from the University of Chicago and surrounding Chicago area community and provided written, informed consent before the start of the study. All experimental procedures were approved by the University of Chicago Institutional Review Board. Participants were paid between $15 and $18, depending on their task performance. Only 28 participants had usable eye-tracking data by the end of Study 2, so to maintain consistency across behavioral and eye-tracking analyses, data from the remaining 32 participants was excluded in the presented results.

Stimuli. Participants in a preliminary Prolific study (N = 80) viewed a new set of 100 unique ShapeStacks images. We categorized images based on their stability judgments: "fall" (< 30% 'STAND' responses), "ambiguous" (35-65% 'STAND' responses), or "stand" (> 65% 'STAND' responses). Based on these categorizations, images were sorted into 4 batches of 25 images each, to be used in the 6 different kinds of experimental blocks in Study 2. Batches A and B contained more *unstable* towers, each with 10 "fall" towers, 8-9 "ambiguous" towers, and 5 "stand" towers. Batches C and D contained more *stable* towers, each with 5 "fall" towers, 8-9 "ambiguous" towers, and 10 "stand" towers. Stimuli were presented in-lab using MATLAB with Psychophysics Toolbox and PsychoPy (Brainard, 1997; Pierce et al., 2019).

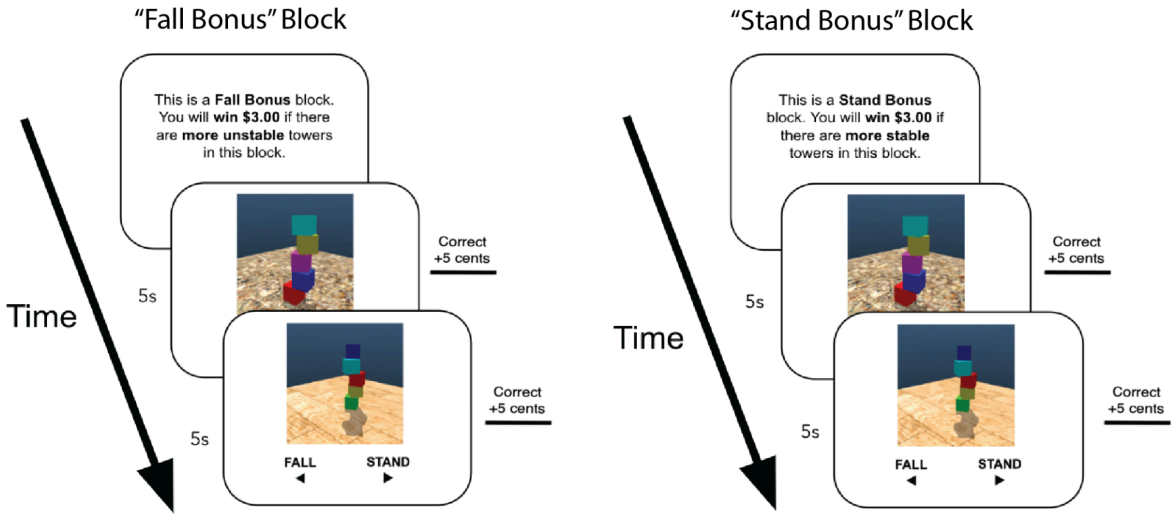

Fig. 7 | Experimental design. a, Participants were presented with different batches of ShapeStacks tower images, one kind with **more unstable images**, and another kind with **more stable images**. At the end of "Fall Bonus" blocks, participants would earn a $3.00 bonus if the block contained more unstable towers than stable towers. At the end of "Stand Bonus" blocks, participants would earn a $3.00 bonus if the block contained more stable towers than unstable towers. If there were more stable images in a "Fall Bonus" block or more unstable images in a "Stand Bonus" block, participants would earn no bonus. Participants won 5 cents for each correct categorization. As in Study 1, participants were given 5 seconds to make their predictions for each tower, by responding with either the 'right' arrow (►) or 'left' arrow (◄) on their keyboard. The reward maximizing strategy was also the same: to earn the most money, participants should ignore the block bonuses and perform the categorizations as accurately as possible.

Behavioral Task. Study 2 removed the social framing (teammates/opponents) while maintaining the motivational manipulation. Participants completed Fall Bonus, Stand Bonus, and Neutral blocks (Fig. 7). In bonus conditions, participants were told they would receive bonuses if blocks contained more stable/unstable images; neutral blocks had no stability-based bonuses. For each of the 3 block types, participants were presented with an image batch containing more unstable images (randomly either Batch A or B), and another containing more stable images (randomly either Batch C or D), for a total of 6 experimental block variations. So that participants were shown every motivation/stability combination, they completed each block twice in random order, for a total of 12 blocks.

Eye-Tracking. Fixation and saccade pattern data was recorded while completing the behavioral task. We tracked both right and left eyes using an EyeLink 1000 Desktop Mount, sampling at 500 Hz. A five-point calibration and validation were run at the beginning of each session. After completing the calibration and validation, participants received instructions about the task. In order to allow participants to rest their eyes and perform the task to the best of their ability, there were four designated breaks, every three blocks.

Results

Behavioral Analysis

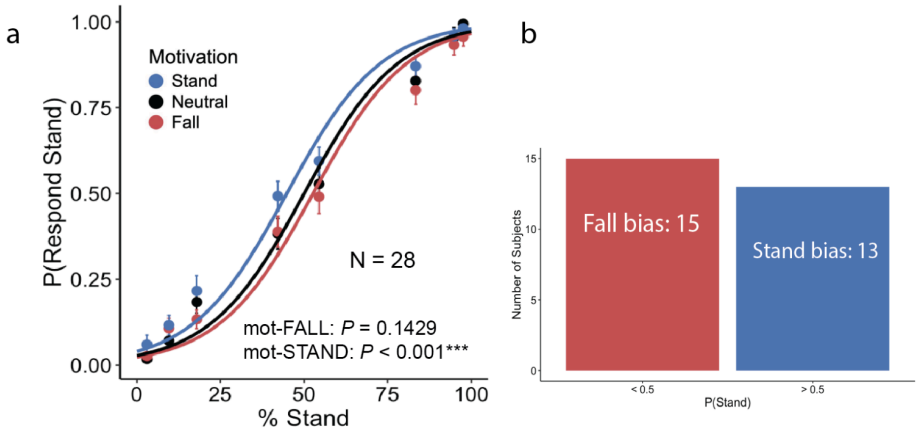

Fig. 8 | Motivation biases physical judgements against an intuitive response bias. Pannels **a** and **b** contain data from 28 participants. **a**, Participants were more likely to report seeing the category they were motivated to see. When participants were motivated to see the tower stand (e.g. mot-STAND, the teammate bet 'STAND' or the opponent bet 'FALL'), their psychometric function was shifted left. Conversely, when participants were motivated to see the tower fall (e.g. mot-FALL, the teammate bet 'FALL' or the opponent bet 'STAND'), their psychometric function was shifted right. **b**, 15 participants had a bias for responding 'FALL', and 13 had a bias for responding 'STAND'.

Motivation biases intuitive physical judgments. As in Study 1, in-lab participants accurately tracked tower stability in neutral conditions (GLME: $z = 58.58$, $p < .001$, $\boldsymbol{b = 0.070}$, SE = 0.0012). Participants were significantly more likely to respond 'STAND' when motivated to see towers stand (GLME: $z = 3.35$, $p < .001$, $\boldsymbol{b = 0.042}$, SE = 0.13) and (insignificantly) less likely when motivated to see towers fall (GLME: $z = -1.47$, $p = .1429$, $\boldsymbol{b = -0.23}$, SE = 0.15) (Fig. 8*a*). In other words, participants were more likely, on average, to report seeing the category they were motivated to see, especially when they were motivated to see the tower stand.

However, in-lab participants also displayed intrinsic response biases (fall-bias = 23/39, stand-bias = 16/39) (Fig. 8*b*), prompting the same subgroup analysis as Study 1.

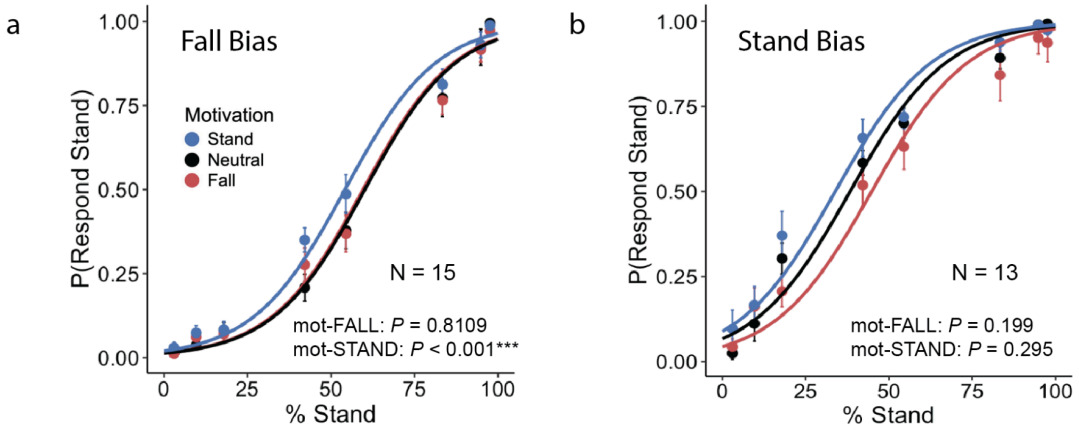

Fig. 9 | Motivation biases physical judgements against an intuitive response bias. a, Fall-bias participants were significantly more likely to be influenced by motivation to see the tower **stand** (e.g. mot-STAND) than motivation to see the tower fall. **b**, Stand-bias participants were (insignificantly) more likely to be influenced by motivation to see the tower **fall** (e.g. mot-FALL) than motivation to see the tower stand.

Motivation biases physical judgements against an intuitive response bias. Both subgroups accurately tracked tower stability in neutral conditions (fall-bias GLME: $z = 36.96$, $p < .001$, $\boldsymbol{b = 0.072}$, SE = 0.0020; stand-bias GLME: $z = 32.89$, $p < .001$, $\boldsymbol{b = 0.069}$, SE = 0.0021). Replicating Study 1's pattern, fall-bias participants were significantly more likely to respond 'STAND' when motivated to see towers stand (GLME: $z = 3.44$, $p < .001$, $\boldsymbol{b = 0.43}$, SE = 0.13) and (insignificantly) more likely to respond 'STAND' when motivated to see towers fall (GLME: $z = 0.24$, $p = .8109$, $\boldsymbol{b = 0.04}$, SE = 0.17) (Fig 9*a*). The pattern of GLME results for the mot-STAND condition in Study 2 differed slightly than Study 1, where stand-bias participants were (insignificantly) *more* likely to respond 'STAND' when motivated to see towers stand (GLME: $z = 1.05$, $p = .295$, $\boldsymbol{b = 0.31}$, SE = 0.29) and (insignificantly) less likely to respond 'STAND' when they were motivated to see the tower fall (GLME: $z = -1.29$, $p = .199$, $\boldsymbol{b = -0.45}$, SE = 0.35) (Fig 9*b*).

Across both studies, motivation consistently biased judgments against participants' intrinsic biases. This "anti-bias" effect was strongest for fall-bias participants with stand-motivation ($p < .001$ in both studies). While the results for stand-bias participants failed to reach significance, a similar pattern was observed (i.e., mot-FALL $p <$ mot-STAND p for both Prolific and in-person participants). We kept these findings in mind throughout the analysis of eye-tracking data.

Eye-Tracking Analysis

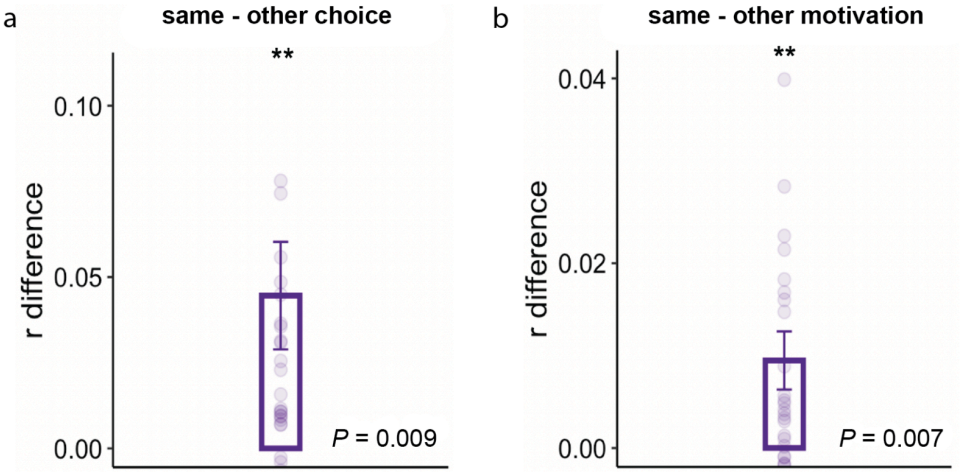

Fig. 10 | Motivation may bias information sampling. a, Fixation patterns were more similar between participants who made the same judgment, suggesting that eye-tracking reflects differences in information sampling. **b**, Fixation patterns were more similar between participants who had the same motivation, suggesting that motivation biases how participants sample information when making judgments.

Motivation biases information sampling. Our primary goal for analyzing fixation data was to be able to quantify *how* people sample information about physical scenes and determine whether motivational contexts affect this process. Therefore, we computed two paired t-tests comparing average fixation data between (1) participants making the same vs. opposite *choices*, and (2) participants with the same vs. opposite *motivation*.

Participants making the same choice showed significantly more similar fixation patterns than those making opposite choices ($t = 2.84$, df = 25, $p = .0088$, 95% CI = 0.012 – 0.077) (Fig. 10*a*). If eye-tracking reflects differences in information sampling, this finding suggests that how people sample information impacts their physical judgments. Participants with the same motivation also showed significantly more similar fixation patterns than those with the opposite motivation ($t = 2.94$, df = 27, $p = .0067$, 95% CI = 0.0036 – 0.020) (Fig. 10*b*). This suggests that motivation influences how people sample the information they use when making judgments.

Exploratory analysis

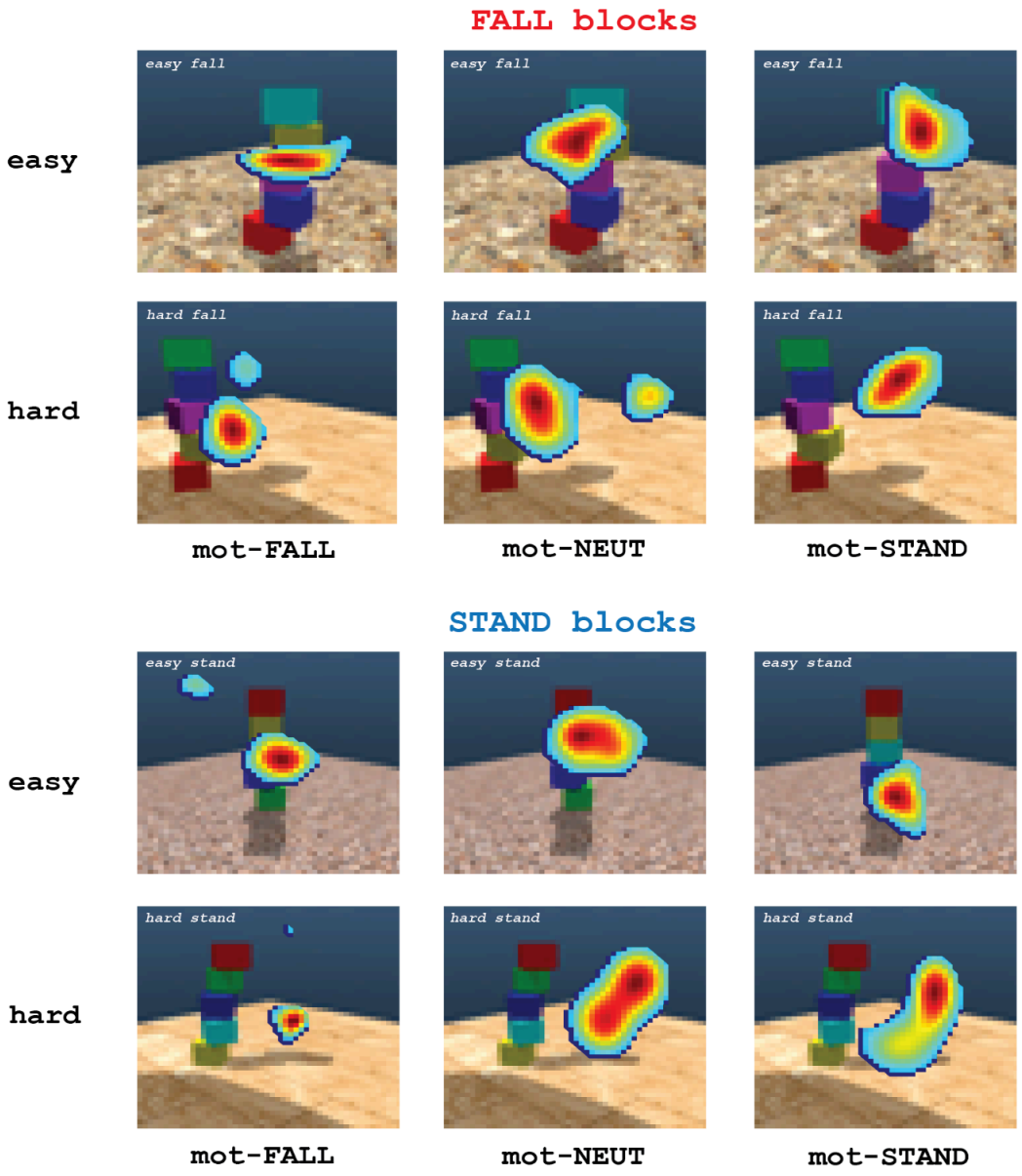

Fig. 11 | Strategies for physical judgments are context-dependent. Averaged (N = 28) heatmap plots of participant fixation points for four different tower images. Heatmaps were produced for unambiguously fall (e.g. *easy fall*), ambiguously fall (e.g. *hard fall*), unambiguously stand (e.g. *easy stand*), and ambiguously stand (e.g. *hard stand*) towers, across each of the motivation conditions (e.g. mot-FALL, mot-NEUT, mot-STAND). For the *easy fall* tower, participants appear to be fixating on a specific block (likely the mechanical failure point) when motivated to see the tower fall. Participants also seem to fixate on the most obvious instability point for the *easy stand* tower in the mot-FALL condition (even though the tower is unambiguously stable). In general, fixations on specific tower blocks seem to occur more frequently for *easy* towers, while saccades in the potential direction of tower collapse seem more common for *hard* towers. This suggests that participants may be **mentally simulating** the possible ways in which the towers might fall in situations with **higher uncertainty**.

Strategies for physical inference may depend on the motivational context. We compared human and AI strategies for physical inference by examining whether participants attend to mechanical instability points, like the Inception v4 classifier. While full statistical analyses are ongoing, preliminary visualizations reveal some salient patterns.

Fig. 11 shows averaged heatmap plots of participant gaze data for four sample towers. As predicted, participants appear to be focusing on the mechanical failure points when they are motivated to see the towers fall. Interestingly, this was the case for both unambiguously unstable (i.e., *easy fall*) and unambiguously stable (i.e., *easy stand*) towers. In other words, even for the *easy stand* tower (when it should have been easy to tell that the tower was stable), the majority of participants appear to have made an attempt to identify a "false" failure point block when they were motivated to perceive the tower as unstable. Conversely, when motivated to see the tower stand, participants seem to focus on a specific block (the "most stable" block, potentially), regardless of the block's objective Fall/Stand ranking. Perhaps participants were attempting to identify a "false" stability point when they were motivated to perceive the tower as stable.

Uncertainty may impact mental simulation. The final question we can begin to address with these preliminary visualizations is the extent to which human physical inference can be instantiated with simulation-based physics engine models like the IPE. An interesting observation from Fig. 11 is that for more ambiguous towers (e.g., *hard fall*, *hard stand*), averaged heatmap plots do not indicate that participants are focusing on any specific block in the tower. Rather, it appears that participants are focusing their attention on a specific *side* of the tower image as a whole. Moreover, the image area covered by the *hard* tower heatmaps seems larger, in general, than the area occupied by the *easy* tower heatmaps. Although we have yet to analyze participant saccades, the existence of these evidently more "dispersed" fixation patterns for the *hard* tower heatmaps might indicate (1) that *mental simulation* is occurring, as assumed by the IPE model (e.g., participants appear to be moving their eyes between two fixation points), and (2) that situational *uncertainty* is affecting the strategies people use to make their judgments (i.e., the "dispersed" fixation patterns are only visible for *hard* towers). These observations align with both the "noisy physics simulator" account of human physical inference and Hamrick et al. (2015)'s SPRT predictions about simulation frequency varying with outcome uncertainty. While these preliminary observations will require more rigorous statistical validation, they provide promising directions for future analyses.

DISCUSSION

The present study found that: (1) participants were more likely to report seeing the physical outcome they were motivated to see, (2) participants making identical judgments showed similar fixation patterns, and (3) participants with the same motivation showed similar fixation patterns. In other words, both participants making similar judgments, as well as participants with the same motivation to see a certain outcome, tended to focus on similar aspects of the towers. Preliminary analyses further suggest that participants fixated on blocks supporting their motivated outcome and relied on mental simulation more frequently in situations with higher uncertainty.

Building on Leong et al. (2019)'s finding that motivation to perceive a desired outcome increases neural activity in category-specific visual areas—suggesting that people literally *see what they want to see*—we found that motivation similarly shapes visual perception in the domain of intuitive physical inference. Across two studies, motivation biased both perceptual judgments and the information sampling processes that precede them. Together, these results

contribute to the "motivated seeing" literature by showing that reward-driven motivation can distort evidence accumulation, leading to inaccurate representations of the world in a *variety* of contexts, from visual categorization to ambiguous physical scenarios. Thus, while people tend to believe that their perceptions are accurate representations of reality, these results underscore the sensitivity of visual processing to external motivational factors such as rewards, risks, and uncertainty. Future work might adopt Leong et al. (2019)'s neuroimaging approach to identify the brain regions that moderate motivational biases in intuitive physical judgments, offering a more comprehensive, computational account of how the drive for reward leads to biased representations of our physical environment.

An unexpected nuance in our analyses suggests that intrinsic bias, too, plays a role in how external motivational factors are integrated into perception. Across both studies, participants exhibited consistent tendencies to respond either 'FALL' or 'STAND,' and these biases interacted systematically with motivational context (e.g., fall-bias participants were significantly more likely to be influenced by stand motivation than fall motivation). Although this "anti-bias" pattern varied somewhat across studies (indeed, this is a limitation of the analyses presented), it is likely that these results are the product of a number of experiment-related factors, such as variability in experimental procedures (e.g., task setup, wording of instructions), experimental setting (e.g., online vs. in-person, size/location of ► and ◄ keyboard buttons), participant sample size, etc. Therefore, future iterations of this work should refine these behavioral measures.

Our preliminary visualizations of averaged heatmaps are promising, both in terms of future research directions and real-world applications. First, participants appeared to fixate on tower blocks that supported the outcome they were motivated to see—mirroring Groth at al. (2018)'s Inception v4 visual classifier, which "learned" to attend to the most unstable block in the tower. Second, fixation patterns suggest that participants may rely more heavily on mental simulation under conditions of greater uncertainty—consistent with Hamrick et al. (2015)'s SPRT predictions about simulation frequency.

Drawing explicit comparisons between human and artificial systems can deepen our understanding of both. For example, it was our understanding of the Inception v4 CNN that informed our subsequent observation that participants focused on aspects of the scene that could serve as evidence for their desired percepts. Conversely, because the IPE seeks to approximate human-like physical reasoning, further analyses of fixation and saccade data may capture more of the nuance in human bias and strategy that can be incorporated into the model's parameters, ultimately leading to more effective AI systems and improved human-AI interaction.

As a closing remark, consider the following anecdote, which showcases the excitement and promise surrounding this area of research, particularly with regard to how human cognition and decision-making may be *enhanced* through the study of AI systems. A recent study found that the development of "superhuman" AIs (like Google DeepMind's program "AlphaGo") appears to have driven human *Go* players to improve their own gameplay (Shin et al., 2023). After analyzing over 5.8 million move decisions by professional *Go* players from 1950 to 2021, the authors were able to determine that humans had, in fact, learned from AI game strategy, allowing them to incorporate more novel moves and make higher-quality decisions. The downstream effects on *Go* players' creativity and decision-making quality were unprecedented, and their impact on human behavior highlights the inherent value in studying any such AI system that models human decision-making.

References

Bates, D., Mächler, M., Bolker, B. & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*, 1-48.

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

Brainard, D.H. (1997). The Psychophysics Toolbox. *Spatial Vision, 10,* 433-436.

Davis, E. & Marcus, G. (2015). The Scope and Limits of Simulation in Cognitive Models. *arXiv preprint, arXiv:1506.04956.*

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), 936-975.

Groth, O., Fuchs, F.B., Posner, I. & Vedaldi, A. (2018). ShapeStacks: Learning Vision-Based Physical Intuition for Generalised Object Stacking. *Proceedings of the European Conference on Computer Vision*. 702-717.

Hegarty, M. (2004). Mechanical reasoning by mental simulation. *Trends in Cognitive Sciences. 8*(6), 280-285.

Johnson-Laird, P.N. (2010). Mental models and human reasoning. *Proceedings of the national academy of sciences. 102*(43), 18243-18250.

Kuznetsova, A., Brockhoff, P. B. & Christensen, R. H. B. (2017). lmerTest: tests in linear mixed effects models. *Journal of Statistical Software*, *82*, 1-26.

Leong, Y.C., Hughes, B.L., Wang, Y., & Zaki, J. (2019). Neurocomputational mechanisms underlying motivated seeing. *Nature Human Behavior, 3,* 962–973.

de Leeuw, J.R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods, 47*, 1-12.

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

Peirce, J., Gray, J.R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., Lindeløv, J.K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavioral Research Methods*, *51*(1), 195-203.

Pylyshyn, Z.W. (2002). Mental Imagery: In search of a theory. *Behavioral and Brain Sciences, 25,* 157-238.

Shin, M., Kim, J., Opheusden, B. V., & Griffiths, T.L. (2023). Superhuman artificial intelligence can improve human decision-making by increasing novelty. *Psychological and Cognitive Sciences, 120*(12),

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.

Zhou, L., Smith, K.A., Tenebaum, J.B., & Gertensberg, T. (2022). Mental Jenga: A counterfactual simulation model of causal judgments about physical support. *Psychological Review, 125*(5), 936-975.