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 of a physical scenario. 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 to perceive the towers as less/more stable by telling them that they would receive a bonus if the presented tower was going to fall/remain standing. However, because the bonus depended on the stability ratings of the tower, and not on participants' responses, the reward-maximizing strategy was to prioritize accuracy over the bonuses. We hypothesized that participants would be motivated to respond that a given tower was stable versus unstable when financially incentivized to make a certain judgment about the tower's stability. 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, fixation patterns were more similar between participants who made the same judgment for a given tower. They were also more similar between participants who were motivated to see the same outcome for a given tower. 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 involve evaluating the extent to which human physical inference can be instantiated with simulation-based physics engine models, as well as exploring the similarities between human strategies for physical inference and those employed by an Inception v4 convolutional neural network architecture. 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. In general, we are able to make quick judgments about the physical relationships between different objects in our environment, but this task becomes harder as a scene becomes more complex. For example, accurately assessing the exact position of each wooden block in 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). With this in mind, understanding the ways in which our motivations, desires, prior knowledge, and/or belief systems play a role in the way we perceive our surroundings is an extremely relevant issue. The current study aims to answer this question: specifically, we seek to investigate 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 computational 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) were able to use a financial incentive 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. Importantly, this result indicated a bias in participants' *neural representation* of the presented image, providing a computational explanation for how reward can (literally) lead to inaccurate representations of the world around us. In other words, participants were actually "seeing" what they wanted to see. Our study builds on this finding in order to provide an account of 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. Indeed, our ability for inferring the physical relationship between objects in our environment is a critical and unique aspect of human intelligence. However, we lack a comprehensive account of the supporting mental processes and computations. For example, years of cognitive science research suggests that our impressive ability for physical intuition depends, at least in part, on the formation of a "mental model" – a mental representation of the objects in a given scenario and their relationships to each other, according to the rules of physics (Johnson-Laird, 2010). However, the detailed nature of these mental models remains only partially understood, as demonstrated by our inability to replicate any genuine human-like computational adaptability in our most modern artificial intelligence (AI) systems (Ludwin-Peery et al, 2021). Thus by investigating this important aspect of human intelligence, researchers hope to not only elucidate our understanding of the relevant cognitive processes in humans, but also bridge the gap between human intelligence and artificial systems and pave the way for more advanced AI capabilities.

Researchers have explored numerous frameworks for explaining the human capacity for physical reasoning, with "mental simulation" emerging as a particularly influential concept. Mental simulation is a cognitive process that involves the internal reconstruction or imagination of a scenario (Davis & Marcus, 2015). Moreover, it is a flexible process that enables people to interpret complex situations, problem-solve, and adapt to their ever-changing environment. 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.

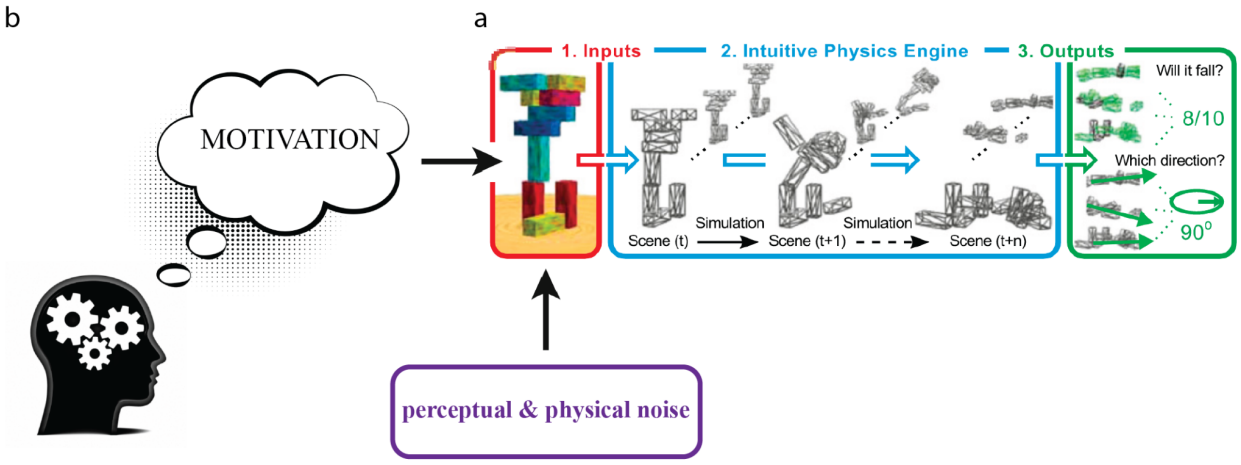

Fig. 1 | The intuitive physics engine (IPE). a, Adapted from Battaglia et al. (2013): It is hypothesized that people reason about physical scenes in the following way: *(1)* As inputs, the model takes an observer's internal reconstruction of a scene, modulated by a set of *noise* parameters (e.g. position, 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.

A recent proposition by Ullman et al. (2017) has introduced an account of mental simulation that can be instantiated using computer models. These models represent the cognitive mechanism underlying human physical judgments with an "Intuitive Physics Engine" (IPE), a system analogous to the physics engines utilized in video games to simulate authentic physical interactions (Fig. 1*a*). In theory, the IPE takes an observer's internal reconstruction of a scene (e.g. their "mental model"), modulated by a set of noise parameters to account for perceptual and physical uncertainties, as inputs. In order to make predictions about how each scene will unfold over time, the IPE runs a series of simulations over a distribution of inputs by applying the relevant laws of Newtonian physics to each. Finally, the IPE outputs the scenario's statistically

most likely outcome – 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 (i.e., the person would perceive the tower as unstable).

The predictions generated by these kinds of simulation-based "physics engine" models have demonstrated a strong correlation with human physical reasoning across an array of tasks, such as the previously mentioned task of assessing the likelihood of a block tower collapsing, as well as predicting the direction in which it will fall (Battaglia et al., 2013). However, further research is needed in order to more precisely define the interplay between visual features, mental simulations, and other heuristic principles in shaping human physical inference (Zhou et al., 2021). For example, Battaglia et al. (2013)'s IPE model assumes that robust and fast physical inference in humans is supported by noisy mental simulations that are *approximate* and *probabilistic*. However, Hamrick et al. (2015) provide evidence that in order to optimally balance speed and accuracy, people will also tend to vary the number of simulations they run, based on their uncertainty in the outcome, according to the sequential probability ratio test (SPRT; also know as the drift-diffusion model). Moreover, on a broader level, the specific decision-making strategies people use may vary based on the goals and characteristics of the specific task at hand – whether there are certain features that constrain the use of simulation, as well as what the mind does when it does *not* use simulation, are both open questions (Ludwin-Peery et al., 2021).

Crucially, the IPE architecture predicts flexible integration with both lower-level perceptuomotor systems and higher-level cognitive systems responsible for planning, action, reasoning, and language (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 (e.g., the observer's internal reconstruction of a scene).

We can explore this effect using eye-tracking fixation and saccade pattern data. For example, when people are motivated to see a particular outcome of a physical scenario, do they 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 the approximate, probabilistic account of mental simulation over the SPRT account?

The most immediate goal of the present study was thus twofold: (1) to extend the "motivated perception" literature to the domain of human physical inference and determine whether people can be motivated to make a particular physical judgment when financially incentivized to make that judgment (Study 1), and (2) to analyze eye-tracking data in order to measure how people sample information about a physical scene in particular motivational contexts (Study 2). Both studies involved a visual categorization task, where we motivated participants to perceive the stability of 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 participant who predicted opposing outcomes.

Another important aim of the present study was to compare intuitive human strategies for physical reasoning with those of AI systems more generally, in order to improve our understanding of the underlying mechanisms in *both* systems. Therefore, the overarching goal for the remainder of our eye-tracking analyses was to compare human strategies for physical inference with those of a relevant deep learning visual classifier.

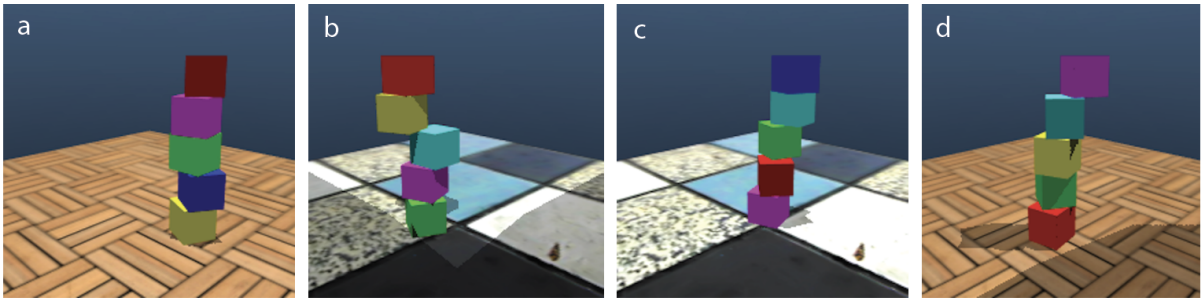

Fig. 2 | Different scenarios from the ShapeStacks dataset. (**a** - **d**) depict initial stack setups: **a**, Stable, rectified tower of cubes. **b**, Unstable tower where multiple objects counterbalance each other. **c**, Stable, but visually challenging scenario due to colors and textures. **d**, 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.

Our study involved presenting participants with a selection of images from the ShapeStacks dataset, a simulation-based dataset derived from training a convolutional neural network (CNN) architecture, Inception v4, to distinguish between stable and unstable structures (Groth et al., 2018). The dataset contains 20,000 images of unstable and stable towers with objective Fall/Stand rankings and a diverse array of structural stability violations (Fig. 2).

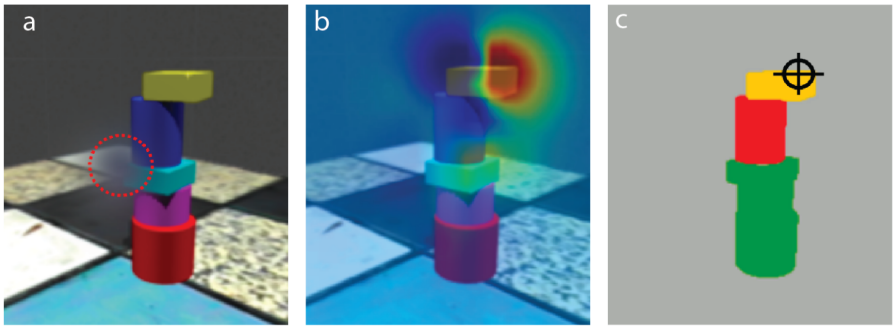

Fig. 3 | ShapeStacks dataset: Inception v4 CNN attention heatmap obtained via occlusion study (Groth et al. 2018). a, A blur is applied to the image. **b**, The increase (red) / decrease (blue) in the predicted stability is shown as a heatmap. **c**, The heatmap is compared to the ground-truth segmentation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

The images also contain annotations of the mechanical failure points of the towers. For example, the crosshairs in Fig. 3*c* highlight the failure point of the tower – the block that, if removed, would render the otherwise stable tower unbalanced and cause it to fall. To determine

whether the network depended on these points to make its judgments, Groth et al. (2018) conducted an "occlusion study," where they examined the impact of partially obscuring an image (e.g. introducing blur) on the neural network's predictions (Fig. 3*a*). If concealing part of an image led the network to change its prediction from unstable to stable, 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*). On average, the center of "attention" of the network was highly correlated with the location of the mechanical failure points, suggesting that through training on a representative dataset, the network had learned to attend to the *points of structural instability* when determining that the tower was unstable (Fig 3*b*, *c*).

Since we know that the ShapeStacks CNN attends to stability violations in order to optimize its decision-making, it is a worthwhile question to ask whether humans, too, might intuitively employ similar strategies to those of the fully-trained CNN. Accordingly, we utilized fixation pattern data to determine whether people direct their attention to the same mechanical failure points identified by the deep learning visual classifier, as well as the motivational context in which they do so. In general, we predicted (1) that participants who were motivated to judge the tower as *stable* would pay *less* attention to the points of structural instability and more to the stability points, and (2) that participants who were motivated to judge the tower as *unstable* would pay *more* attention to the instability points, perhaps in order to inform their mental reconstruction of the scenario and 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 block tower images from the ShapeStacks dataset. The towers varied both in terms of their objective stability ratings as well as in terms of the difficulty with which subjects were expected to have while making a judgment – 10 of the 40 towers presented were unambiguously stable (e.g. “easy stand”), another 10 towers were unambiguously unstable (e.g. “easy fall”), another 10 towers were ambiguously stable (e.g. “hard stand”), and the last 10 towers were ambiguously unstable (e.g. “hard fall”). To present the stimuli remotely via a web browser, the experiment was coded in JavaScript and hosted on the online data storage platform Cognition (www.cognition.run) [2022]. A significant portion of the experiment code was written using JsPsych, a JavaScript library for creating behavioral experiments in a web browser (de Leeuw, 2015).

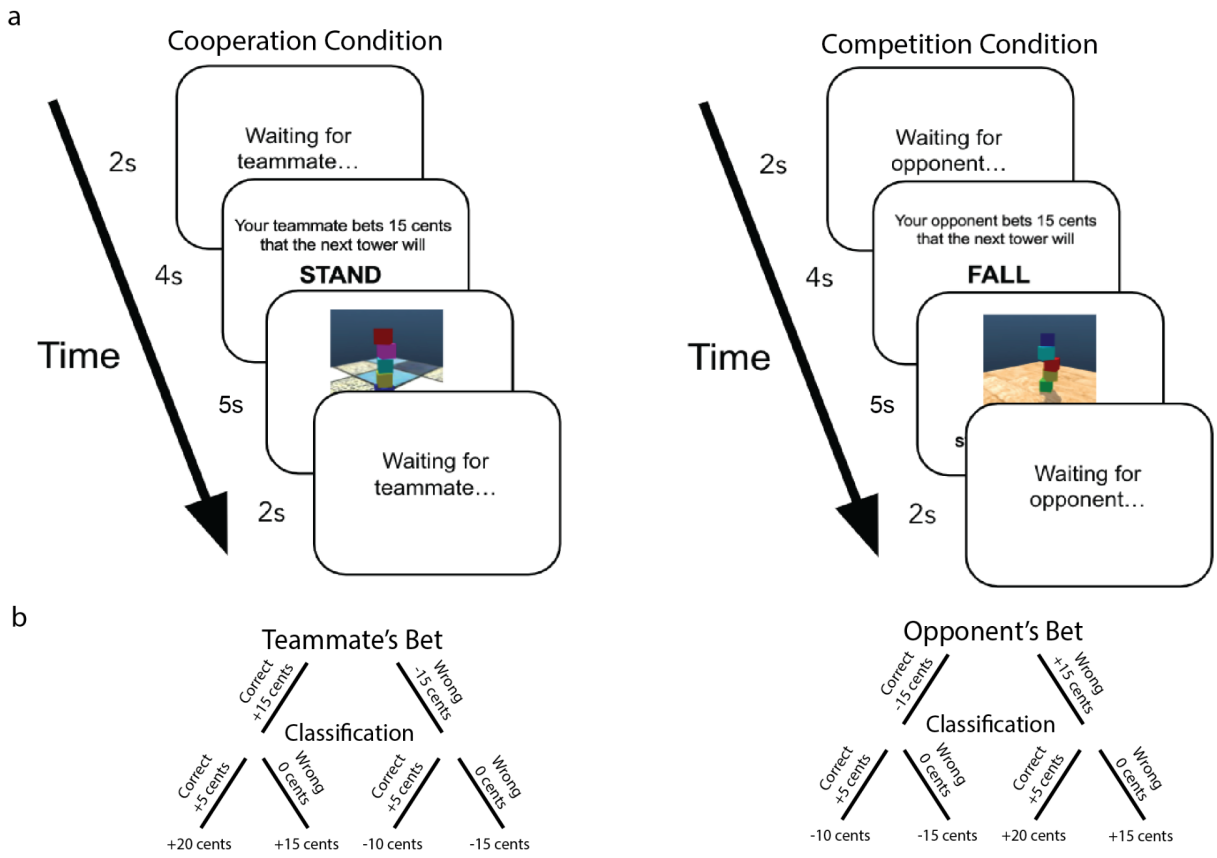

Fig. 4 | Experimental design. a, Participants were presented with ShapeStacks tower images. In the cooperation condition, a teammate first makes a bet about whether the tower will fall or stand. Participants are then presented the image and have to similarly determine whether the tower will fall or stand. 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.

Behavioral Task. There were three experimental blocks for three conditions ('neutral', 'cooperation', and 'competition'), and blocks consisted of 40 trials. In each trial of the neutral condition, participants were presented with the 40 tower images and were given 5 seconds to make a prediction of either 'FALL' or 'STAND' for each tower, by responding with either the 'right' arrow (►) or 'left' arrow (◄) on their keyboard. For the cooperation and competition

conditions (Fig. 4*a*), participants were told that they would be performing a visual categorization task with a 'teammate' (cooperation) and an 'opponent' (competition). In each of these two conditions, the participant's teammate/opponent would make a bet of either 'FALL' or 'STAND' on the upcoming image (taken from the original set of 40 from the neutral condition). Participants were then presented with the same image and given 5 seconds to predict either 'FALL' or 'STAND'. In the cooperation condition, if the teammate's bet was correct, both the teammate and the participant won a monetary bonus. If the teammate's bet was wrong, both the teammate and the participant would lose money. The payoff structure of the competition condition was identical to the cooperation condition, except that if the opponent's bet was correct, the opponent would earn a bonus and participants would lose money. If the opponent's bet was wrong, the opponent would lose money and participants would earn a bonus (Fig. 4*b*).

We included *two* conditions in the task in order to control for the effect of semantic priming – for example, when the teammate bets 'STAND', participants might be more likely to *respond* 'STAND' simply because they were semantically primed by having just seen the word 'stand', and not because they were motivated to *see* the tower stand. Motivating participants to respond both consistently and inconsistently with the presented bet thus allowed us to directly test this competing account. Finally, it is also important to note that the outcome of the teammate's and opponent's bets were determined by the objective Fall/Stand rankings of the presented images, and not by participants' subjective categorizations. Therefore, in order to earn the most money, participants should have ignored the bets and made their categorizations accurately (Fig. 4*b*).

Psychometric functions. To model the behavioral data (from both Study 1 and Study 2) and estimate the impact of motivation on participants' categorizations, we employed a

Generalized Linear Mixed Effects (GLME) model. The first kind of model was fit to data from all conditions (e.g. cooperation, competition, neutral) and participants. The other kind of model was fit to data from all conditions, but for two particular participant subgroups in our analyses – one group with an inherent bias to respond 'FALL', and the other with an inherent bias to respond 'STAND'.

Each model incorporated the percent likelihood of a participant responding 'STAND' (where response 0 = 'FALL' and response 1 = 'STAND') as a covariate. To account for random variability across participants, models also included random intercepts and random slopes for the effects of the motivation condition. Models were estimated using the glmer function in the lme4 package in R, with *P* values calculated from t-tests with Satterthwaite approximation for the degrees of freedom as implemented in the lmerTest package (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

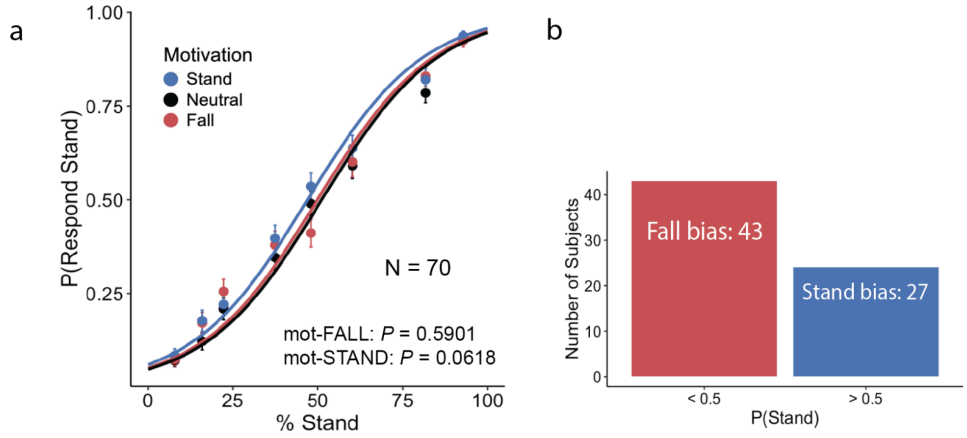

Fig. 5 | Motivation biases intuitive physical judgements. Pannels **a** and **b** contain data from 70 participants. **a**, Participants were more likely to report seeing the category they were motivated to see. When participants were motivated to see the tower stand (e.g. mot-STAND, the teammate bet '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. Not surprisingly, when estimating the psychometric functions for situations where participants had not been motivated to see a particular outcome (e.g. the neutral condition), we found that participants were significantly more likely to respond 'STAND' (GLME: $z = 45.5$, $P < 0.001$, $\boldsymbol{b} = \mathbf{0.059}$, SE = 0.0013). To examine the effect of motivation across all participants, we estimated separate psychometric functions for situations where participants had been motivated to see the tower stand (i.e. either the teammate bet 'STAND' or the opponent bet 'FALL') and 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: $z = 1.87$, $P = 0.0618$, $\boldsymbol{b} = \mathbf{0.25}$, SE = 0.13) and

(insignificantly) more likely to respond 'STAND' when they were motivated to see the tower fall (GLME: $z = 0.54$, $P = 0.5901$, $\boldsymbol{b} = \mathbf{0.074}$, SE = 0.14).

Although we predicted that participants would respond 'STAND' more frequently when motivated to see the tower stand, and conversely, that participants would respond 'FALL' more frequently when motivated to see the tower fall, the preceding analysis produced this pattern only for conditions where participants were motivated to see the towers stand. However, across the entire participant group in Study 1, we did find that individual participants exhibited a bias towards one of the responses (e.g., they had a general tendency to indicate towers as stable or unstable). Specifically, 43/70 participants exhibited a bias for responding 'STAND' in the neutral condition (where they responded 'STAND' for more than 50% of the images), and 27/70 participants exhibited a bias for responding 'FALL' in the neutral condition (where they responded 'FALL' for more than 50% of the images) (Fig. 5*b*). Therefore, we decided to examine the effect of motivation for each of these two participant subgroups separately.

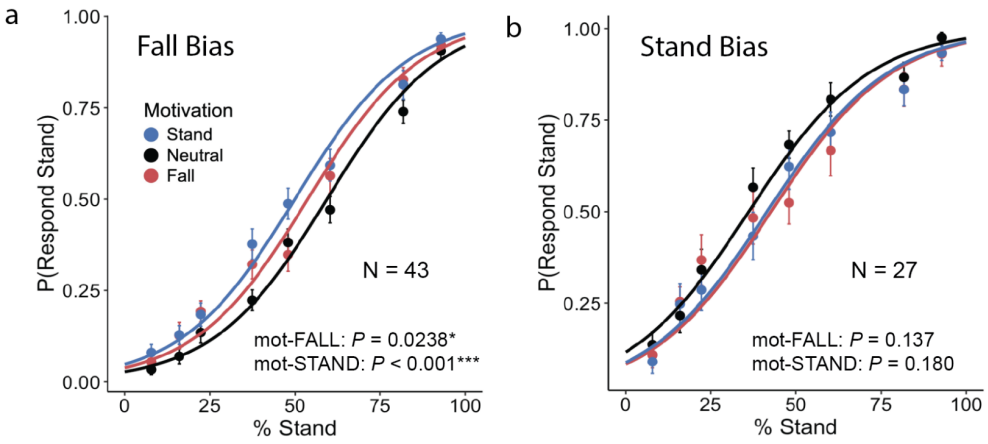

Fig. 6 | Motivation biases physical judgements against an intuitive response bias. a, *Fall bias* participants were more significantly likely to be influenced by motivation to see the tower **stand** (e.g. mot-STAND) than motivation to see the tower fall. **b**, *Stand bias* participants were (insignificantly) 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. As expected, both *fall bias* (*fb*) participants and *stand bias* (*sb*) participants were significantly more likely to respond 'STAND' for blocks with an objective Stand ranking in the neutral condition (*fb* GLME: $z = 37.09$, $P < 0.001$, $\boldsymbol{b = 0.060}$, SE = 0.0016 / *sb* GLME: $z = 26.1$, $P < 0.001$, $\boldsymbol{b = 0.055}$, SE = 0.0021).

To determine whether motivation differentially influenced participants' judgments based on their inherent response biases, we calculated separate psychometric functions for *fall bias* and *stand bias* participants (Fig. 6). We found that *fall bias* participants were significantly more likely to respond 'STAND' when they were motivated to see the tower stand (*fb* GLME: $z = 4.05$, $P < 0.001$, $\boldsymbol{b = 0.56}$, SE = 0.14) and (insignificantly) more likely to respond 'STAND' when they were motivated to see the tower fall (*fb* GLME: $z = 2.26$, $P = 0.024$, $\boldsymbol{b = 0.34}$, SE = 0.15) (Fig 6*a*). *Stand bias* participants were (insignificantly) less likely to respond 'STAND' when they were motivated to see the tower stand (*sb* GLME: $z = -1.34$, $P = 0.180$, $\boldsymbol{b = -0.30}$, SE = 0.23) and (insignificantly) less likely to respond 'STAND' when they were motivated to see the tower fall (*sb* GLME: $z = -1.37$, $P = 0.137$, $\boldsymbol{b = -0.37}$, 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*, more frequently that they would have without the influence of motivation. 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 were presented with 100 unique images from the ShapeStacks dataset. Before we administered Study 2 to participants, we conducted a quick Prolific study (N = 80), where we had participants make basic stability judgments (*without* the influence of motivation) for the new set of 100 images. Based on their responses, we categorized each of the images as either “fall” (where < 30% of participants responded ‘STAND’), “ambiguous” (where 35-65% of participants responded ‘STAND’), or “stand” (where > 65% of participants responded ‘STAND’). 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. For the in-lab experimental task, stimuli were presented using MATLAB software (MathWorks) and the Psychophysics Toolbox, and the experiment code itself was integrated with code written using the PsychoPy Python library (Brainard, 1997; Pierce et al., 2019).

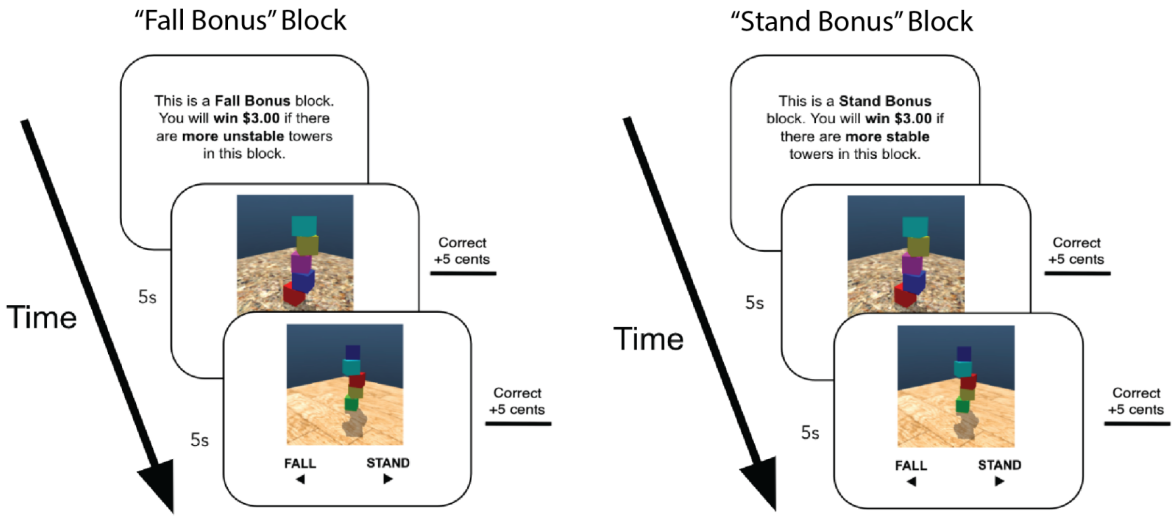

Fig. 7 | Experimental design. a, Participants were presented with different batches of ShapeStacks tower images, one kind with **more unstable images**, and another kind with **more stable images**. At the end of "Fall Bonus" blocks, participants would earn a $3.00 bonus if the block contained more unstable 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. In Study 2, we removed the "social" element of the teammates, opponents, and bets from Study 1. Participants were still motivated to either see the towers fall or stand, or given no motivation at all. Fig. 7 shows the progression of "Fall Bonus" and "Stand Bonus" blocks. At the beginning of each block, we motivated participants to perceive the towers

as more/less stable by informing them that they would receive a bonus if the upcoming block contained more stable images or more unstable images. The "Neutral Bonus" block setup was identical, except participants were told upfront that they would not receive a bonus based on the proportion of stable or unstable images in the block. 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 different kinds of experimental blocks. So that participants were shown with every motivation/stability combination, they completed each block twice, in random order, for a total of 12 blocks overall.

Eye-Tracking. Participants' fixation and saccade pattern data were 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

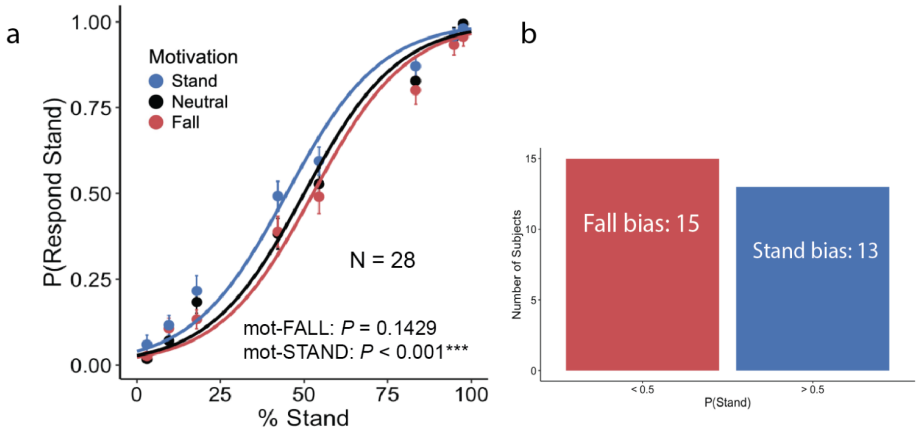

Fig. 8 | Motivation biases physical judgements against an intuitive response bias. Pannels **a** and **b** contain data from 28 participants. **a**, Participants were more likely to report seeing the category they were motivated to see. When participants were motivated to see the tower stand (e.g. mot-STAND, 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 were significantly more likely to respond 'STAND' in the neutral condition for blocks with an objective Stand rating (GLME: $z = 58.58$, $P < 0.001$, $\boldsymbol{b = 0.070}$, SE = 0.0012). Moreover, participants were significantly more likely to respond 'STAND' when they were motivated to see the tower stand (GLME: $z = 3.35$, $P < 0.001$, $\boldsymbol{b = 0.042}$, SE = 0.13) and (insignificantly) less likely to respond 'STAND' when they were motivated to see the tower fall (GLME: $z = -1.47$, $P = 0.1429$, $\boldsymbol{b = -0.23}$, 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.

While this pattern of results does provide support for our initial hypothesis, we found that in-lab participants, too, displayed intrinsic response biases (*fall bias* = 23/39, *stand bias* = 16/39)

(Fig. 8*b*). Therefore, in order to examine the effect of motivation for each of these two participant subgroups separately, we conducted the same follow-up analysis for in-lab participants as we did for Prolific participants.

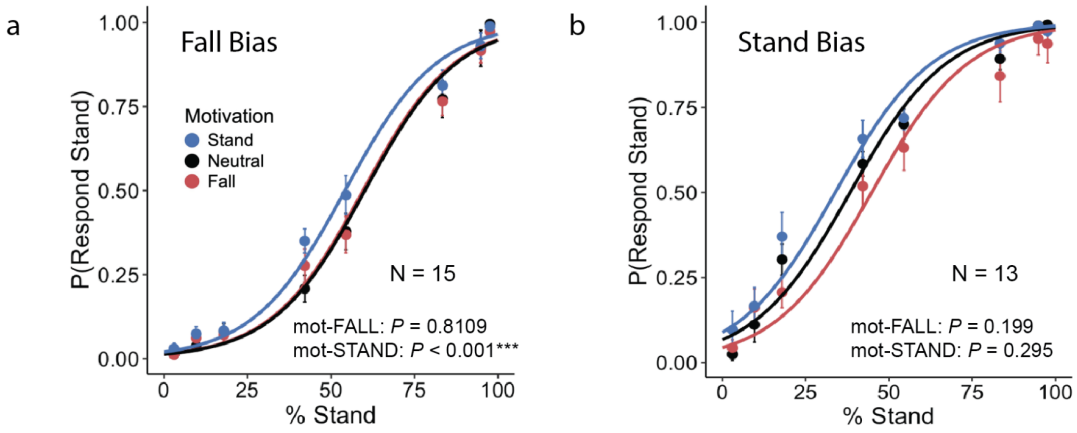

Fig. 9 | Motivation biases physical judgements against an intuitive response bias. a, *Fall bias* participants were significantly more likely to be influenced by motivation to see the tower **stand** (e.g. mot-STAND) than motivation to see the tower fall. **b**, *Stand bias* participants were (insignificantly) 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. As expected, when decomposing the original GLME model by participant response bias, both *fall bias* and *stand bias* participants were significantly more likely to respond 'STAND' for blocks with an objective Stand ranking in the neutral condition (*fb* GLME: $z = 36.96$, $P < 0.001$, $\boldsymbol{b} = \mathbf{0.072}$, SE = 0.0020 / *sb* GLME: $z = 32.89$, $P < 0.001$, $\boldsymbol{b} = \mathbf{0.069}$, SE = 0.0021).

As in Study 1, *fall bias* participants were significantly more likely to respond 'STAND' when they were motivated to see the tower stand (*fb* GLME: $z = 3.44$, $P < 0.001$, $\boldsymbol{b} = \mathbf{0.43}$, SE = 0.13) and (insignificantly) more likely to respond 'STAND' when they were motivated to see the tower fall (*fb* GLME: $z = 0.24$, $P = 0.8109$, $\boldsymbol{b} = \mathbf{0.04}$, SE = 0.17) (Fig 9*a*). The pattern of GLME results for the mot-STAND condition in Study 2 differed slightly than the pattern in Study 1, where *stand bias* participants were (insignificantly) ***more*** likely to respond 'STAND' when they

were motivated to see the tower stand (*sb* GLME: $z = 1.05$, $P = 0.295$, $\boldsymbol{b = 0.31}$, SE = 0.29) and (insignificantly) less likely to respond 'STAND' when they were motivated to see the tower fall (*sb* GLME: $z = -1.29$, $P = 0.199$, $\boldsymbol{b = -0.45}$, SE = 0.35) (Fig 9*b*).

Dividing the behavioral data from both Study 1 and Study 2 into groups based on Fall/Stand biases suggests that motivation may have had the effect of biasing physical judgments against participants' intrinsic biases. Importantly, this "anti-bias" effect was the strongest for *fall bias* participants who had been motivated to see the tower stand (e.g., mot-STAND $P < 0.001$*** for both Prolific and in-person participants). While the results for *stand bias* participants failed to reach significance, a similar pattern to the *fall bias* participants was observed (e.g. 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

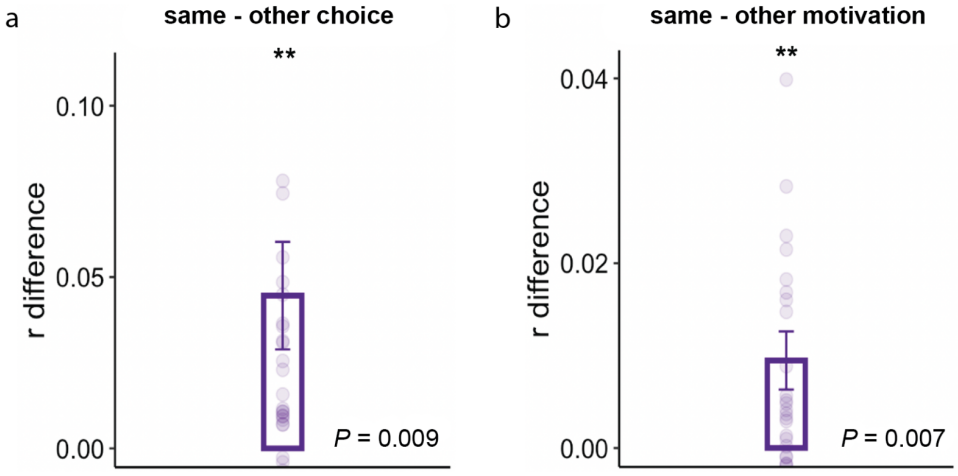

Fig. 10 | Motivation may bias information sampling. a, Fixation patterns were more similar between participants who made the same judgment, suggesting that eye-tracking reflects differences in information sampling. **b**, Fixation patterns were more similar between participants who had the same motivation, 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 a physical scene, as well as to be able to determine whether there are certain (motivational) contexts that affect this process. Therefore, we computed two paired t-tests – one to compare the average fixation data of all participants who made the **same choice** with that of participants who made the **opposite choice**, and one to compare the average fixation data of all participants who had the **same motivation** with that of participants who had the **opposite motivation**.

Results of the first t-test indicated that there was a statistically significant difference between the fixation patterns of participants who made the same choice and those who made the opposite choice ($t = 2.84$, df = 25, $P = 0.0088$, 95% CI = 0.012 - 0.077) (Fig. 10*a*). In other words, fixation patterns of participants who had the same motivation were *more similar to each other* than fixation patterns of participants who had the opposite motivation. If eye-tracking reflects differences in sampling, this finding that participants who made identical choices tended to focus on more similar aspects of the towers suggests that *how* people sample information impacts their physical judgments.

Results of the second t-test revealed that there was also a statistically significant difference between the fixation patterns of participants who had the same motivation and those who had the opposite motivation ($t = 2.94$, df = 27, $P = 0.0067$, 95% CI = 0.0036 - 0.020) (Fig. 10*b*). In other words, fixation patterns of participants who had the same motivation were *more similar to each other* than fixation patterns of participants who had the opposite motivation. In the context of the results of the first t-test, this finding that participants with the same *perceptual goals* tended to focus on more similar aspects of the towers suggests that motivation may influence how people sample the information they use when making judgments.

Exploratory analysis

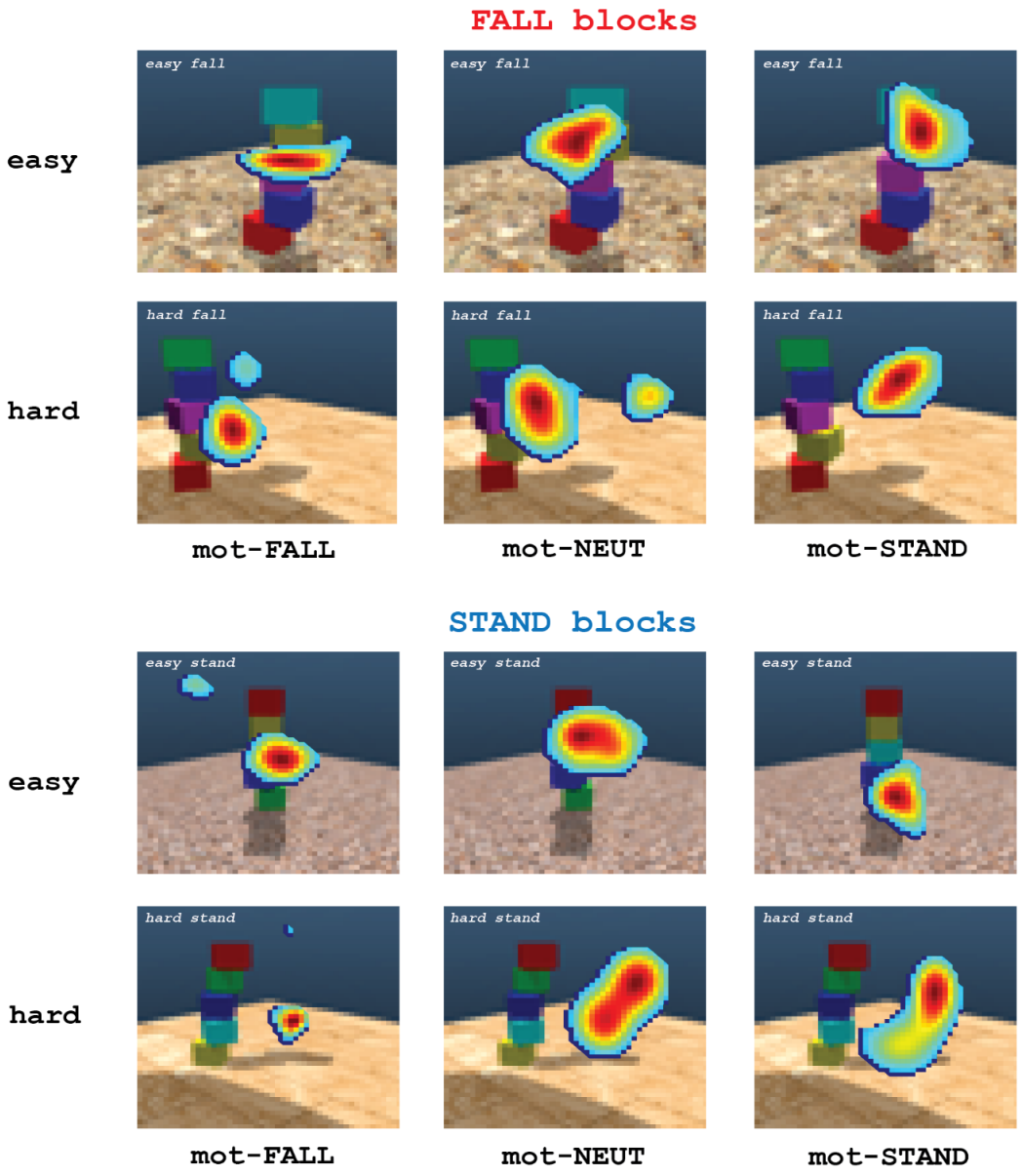

Fig. 11 | Strategies for physical judgments are context-dependent. Averaged (N = 28) heatmap plots of participant fixation points for four different tower images. Heatmaps were produced for unambiguously fall (e.g. *easy fall*), ambiguously fall (e.g. *hard fall*), unambiguously stand (e.g. *easy stand*), 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. One way we can make comparisons between human and AI strategies for physical inference is by utilizing fixation data to determine whether humans and computer models use the same kinds of information to inform their decisions. Since we know that the Inception v4 visual classifier "learned" to optimize its decision-making by attending to stability violations in the towers, we similarly wanted to assess whether participants were reliably directing their attention to areas of mechanical instability. These analyses are still in the early stages, and future statistical work will focus on quantifying the relationships between the averaged participant heatmaps and the network's "attention maps" for each ShapeStacks image used in our study. However, we have competed a preliminary visualization of participant fixation patterns, which we can use to begin thinking about this and related questions.

Fig. 11 depicts the averaged heatmap plots of participant gaze data for four sample ShapeStacks images that were presented to participants during Study 2. As we predicted, participants *do* 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. The inverse pattern emerged for *easy* towers for participants who were motivated to see the tower stand. These participants also 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 (e.g., the "dispersed" fixation patterns are only visible for *hard* towers). So, not only do these observations coincide with the "noisy physics simulator" account of human physical inference, but they also support Hamrick et al. (2015)'s prediction that people should vary the number of simulations they run based on their uncertainty in the outcome, according to the SPRT. While we cannot make any substantial claims without a more rigorous statistical analysis of our entire stimulus set, these preliminary observations are promising and will inform the direction of our future analyses.

DISCUSSION

The present study had three main findings: (1) participants were more likely to report seeing the outcome of an ambiguous physical scenario they were motivated to see, (2) participant fixation patterns were more similar to fixation patterns of other participants who made the same choice, and (3) participant fixation patterns were more similar to fixation patterns of other participants who had the same motivation. 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. Additionally, there were two important observations from the preliminary exploratory analyses: (1) participants appeared to fixate on the tower blocks that supported the outcome they were motivated to see, and (2) participants appeared to rely on mental simulation more frequently in situations with higher uncertainty than in situations with lower uncertainty.

Building from Leong et al. (2019)'s finding that motivation to see a particular visual percept increases neural activity specific to the motivationally relevant visual category (i.e., participants were literally "seeing" what they wanted to see), we found in our first study that motivation could similarly influence visual perception within the domain of intuitive physical inference. Findings from our second study indicate that motivation biases the information sampling process preceding judgment. Taken together, these results contribute to the "motivated seeing" literature by suggesting that the drive for reward can bias the accumulation of evidence, leading to inaccurate representations of the world in a *variety* of visual contexts, such as visual categorization and ambiguous physical scenarios. Thus, while people tend to believe that their perceptions are accurate representations of the world, it is evident that visual perception is highly susceptible to external motivational factors such as rewards, risks, and uncertainty. Future work

might also adopt Leong et al. (2019)'s neuroimaging approach to identify the brain regions that moderate motivational biases in intuitive physical judgments. This has the potential to offer a more comprehensive, *computational* account of how the drive for reward leads to biased representations of our physical environment.

Moreover, an unexpected nuance in our GLME analyses suggests that intrinsic bias, too, plays a role in how these external motivational factors are integrated into perception. Participants in both of our studies displayed strong biases for responding either 'FALL' or 'STAND', and in general (most obviously for *fall bias* participants), their biases were related to the kind of motivation that was most likely have an influence on their judgments (e.g. *fall bias* participants were significantly more likely to be influenced by **stand** motivation than fall motivation). Although there were slight discrepancies in this "anti-bias" pattern of GLME regression results between Studies 1 and 2 (indeed, this is a limitation of the analyses presented), it is likely that these results are the product of a number 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, our behavioral results will need further definition in future iterations of studies for this project, and future research should also account for the fact that motivation appears to have encouraged participants to respond in opposition to their inherent response biases, rather than merely motivating them to report the category they were inclined to see, our behavioral results need further definition in future iterations of studies for this project.

Finally, our observations from visualizing the averaged heatmap plots for participant data are promising, both in terms of future research directions and real-world applications. Our first main takeaway is that participants may fixate on the particular tower blocks that support the outcome they are motivated to see (much like Groth at al. (2018)'s Inception v4 visual classifier that "learned" to pay attention to the most unstable block in the tower). Secondly, participants may rely on mental simulation more frequently in situations with higher uncertainty than in situations with lower uncertainty (much like the "noisy physics simulator" that varies the number of simulations in accordance with the SPRT).

Efforts to make such comparisons between the supporting processes in humans and artificial systems can improve our understanding of both systems, and the results of the present study indicate how future research might accomplish these and other advances. For example, it was our understanding of the Inception v4 CNN that informed our subsequent observation that participants were focusing on aspects of the scene that could serve as evidence for the percept they were motivated to see. Conversely, since the IPE is a computational model that aims to approximate human-like physical reasoning, further exploratory analyses of fixation and saccade data from our study 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.