A stochastic world model on gravity for stability inference

Authors: Taicheng Huang[1], Jia Liu[1*]

[1]Department of Psychology and Tsinghua Laboratory of Brain & Intelligence, Tsinghua University, Beijing, China.

* Correspondence to: liujiathu@tsinghua.edu.cn (J. Liu).

Abstract

The fact that objects without proper support will fall to the ground is not only a natural phenomenon, but also common sense in mind. Previous studies suggest that humans may infer objects' stability through a world model that performs mental simulations with *a priori* knowledge of gravity acting upon the objects. Here we measured participants' sensitivity to gravity's direction, the most critical parameter of gravity in stability inference, to investigate how the world model works. We found that the world model was not a faithful replica of Newton's law of gravity but rather encoded gravity's direction as a Gaussian distribution, with the vertical direction as the maximum likelihood. The world model with this stochastic feature fit nicely with participants' subjective sense of objects' stability and explained the illusion that taller objects are perceived as more likely to fall. Furthermore, a computational model with reinforcement learning revealed that the stochastic feature likely originated from agent-environment interaction, and computer simulations illustrated the ecological advantage of the stochastic over deterministic representation of gravity's direction in balancing accuracy and speed for efficient stability inference. In summary, the stochastic world model on gravity provides an example of how *a priori* knowledge of the physical world is implemented in the brain that helps humans operate flexibly in open-ended environments.

Introduction

About two thousand years ago, Confucius warned his disciples that a wise man should not stand next to a collapsing wall. We, wise or not, can easily judge whether a wall is stable or collapsing in a fraction of a second (Battaglia et al., 2013; Kubricht et al., 2017; McCloskey, 1983). This astonishing performance is unlikely to have been achieved by previous visual experience alone. Taking a stack consisting of ten blocks as an example (Fig. 1), we can quickly report its stability with a satisfactory accuracy of 70% on average (Bear et al., 2021; Zhang et al., 2016), but the universal cardinality of possible configurations is at least 3.72×10^{19} (Extended Data Fig. 1), which is much larger than the total number of sand grains on Earth (est. 7.5×10^{18}) (Blatner, 2013). Contrary to this intuition, four-month-old infants, who have a little visual experience of the physical world, expect a box to fall if it loses contact with a support platform (Baillargeon, 1994, 2004). Our minds may therefore have devised a mechanism that differs from the widely used discriminative approach in artificial neural networks, which relies on the extensive visual experience of objects and feedback about their stability (Bear et al., 2021; Li et al., 2016; Zhang et al., 2016).

Indeed, both behavioral and neuroimaging studies have suggested that humans possess *a priori* knowledge of Newton's law of physics in the mind. For example, infants as young as seven months expect a downward moving object to accelerate and an upward moving object to decelerate (Friedman, 2002; Kim & Spelke, 1999), and adults can estimate the remaining time to catch a moving ball (McIntyre et al., 2001; Zago & Lacquaniti, 2005) even in the absence of visual information (Lacquaniti & Maioli, 1989; Zago et al., 2009). Further fMRI studies have revealed the parieto-insular vestibular cortex in the brain as the neural basis for gravity-based stability inference, suggesting that this knowledge is encapsulated as a cognitive module (Fischer et al., 2016; Indovina et al., 2005; Pramod et al., 2022). Accordingly, our brain is proposed as a set of generative machines that actively predict future events of the ever-changing physical world through mental simulation with *a priori* knowledge acting upon the world (Battaglia et al., 2013; Hegarty, 2004; Huang & Rao, 2011; Tenenbaum et al., 2011; Ullman et al., 2017). For this reason, the generative machine is also called the world model (Land, 2014; Tenenbaum et al., 2011).

Recently, the idea of the world model has become popular to explain the predictive nature of the brain (Friston et al., 2021) and to improve the generality and robustness of the artificial neural networks (Matsuo et al., 2022). However, how *a priori* knowledge is implemented in the world model remains to be determined. A widely adopted but not rigorously tested assumption is that the world model in the brain is a faithful replica of the physical laws of the world (Allen et al., 2020; Battaglia et al., 2013; Lake et al., 2017; Zhou et al., 2022). For example, the direction of gravity encoded in the world model, which is the most critical parameter for stability inference, is assumed to be straightly downward, the same as the direction of gravity in the physical world. Alternatively, there is a consensus that the brain actively correlates, integrates, and comprehends the data from sensory organs (e.g., electromagnetic waves from the eyes) and adds meaning to them (i.e., color). Therefore, the representation of the world in the brain may not be the same as reality. Here, we investigated these two alternative hypotheses for the construction of the world model in the brain by examining how gravity's direction was represented in the world model when participants judged the stability of objects.

To do this, we measured participants' sensitivity to gravity's direction in a stability inference task (Battaglia et al., 2013) and found that gravity's direction was encoded in a Gaussian distribution, with the vertical direction as the maximum likelihood. This stochastic parameter was then built into the world model to simulate the displacement of blocks in a stack under the force of gravity, and the simulation result fits nicely with participants' judgment of stacks' stability and explained the daily illusion that taller objects are perceived as more like to fall. A computational model with a reinforcement learning algorithm was devised to reveal its origin through interactions with the physical world. Finally, we explored the ecological advantage of the stochastic feature of the world model.

Results

The direction of gravity in the world model

The direction of gravity is perpendicular to the ground surface. Here, we first tested humans' sensitivity to gravity's direction to investigate how faithfully our gravity is represented in the world model compared to gravity in the physical world. To do this, we used Pybullet (Coumans & Bai, 2016), a forward physics simulator, to manipulate gravity's direction. Then, we asked the participants to judge whether the collapse trajectories of unstable stacks were normal (Fig 1a, Supplementary Movie S1). The direction of simulated gravity was measured by a parameter pair (θ, φ) (Fig 1b), which determines the deviation of the direction of simulated gravity from the direction of gravity in the physical world. Specifically, θ is the vertical component of the direction that affects the degree of collapse, and φ is the horizontal component that determines the orientation of collapse. We collected participants' judgment of the normality of collapse trajectories while varying θ from 0 to 45° and φ from 0° to 360° across the force space, and the normality ratio of the judgment for each angle pair was used to index participants' sensitivity to gravity's direction (Fig 1c). As expected, when θ is equal to 0 (i.e., the direction of the simulated gravity is the direction of the natural gravity), the participants were likely to report that the collapse trajectory was normal (accuracy: 91.0%, STD: 8.0%). Then, the critical question is how participants' subjective sense about the normality of collapse trajectories changes as a function of θ. If our world model on gravity is a faithful replica of the physical reality, we should expect the immediate detection of abnormality when θ is away from 0.

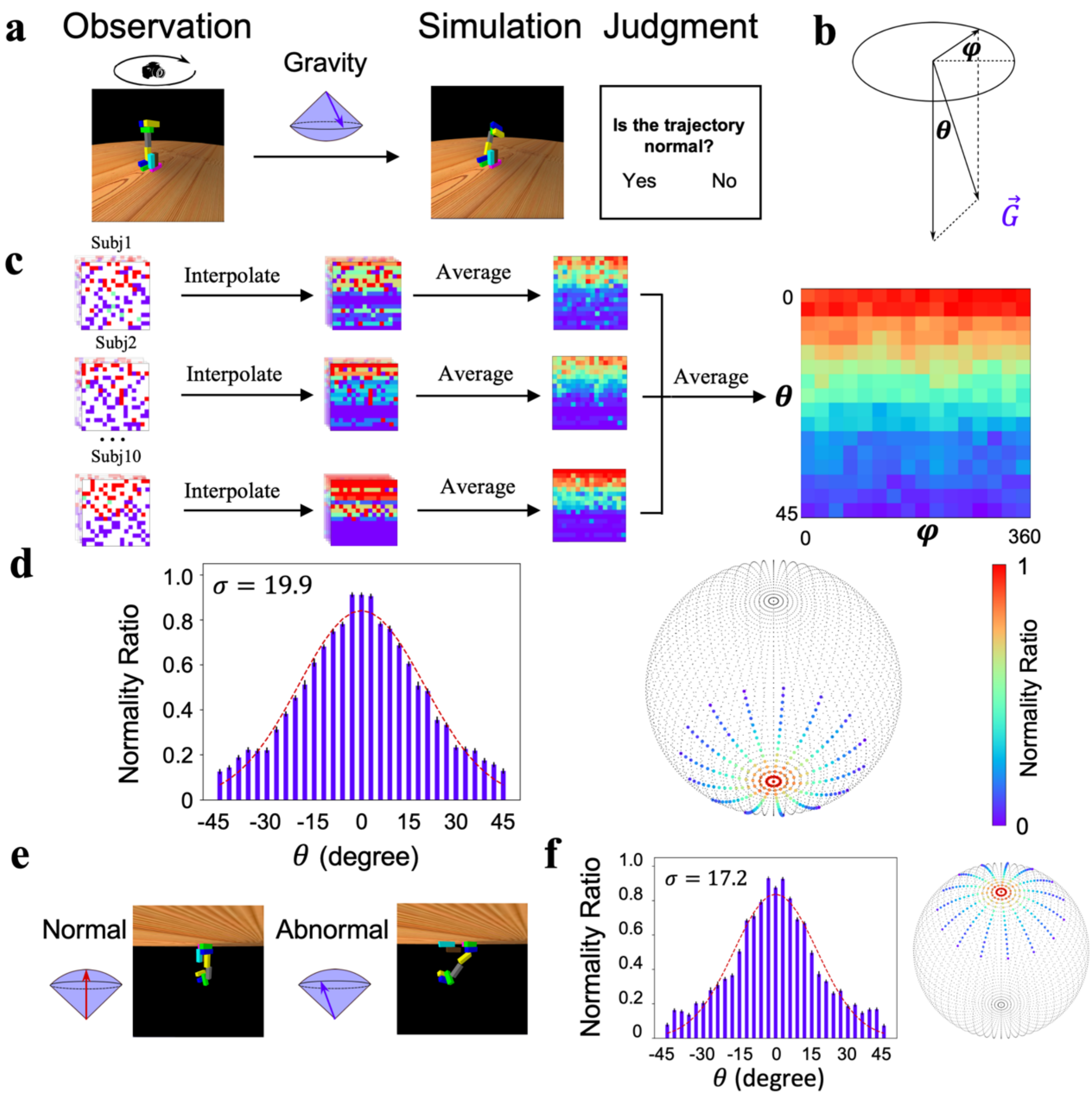


Fig 1. Gravity's direction in the world model. a) The design of the behavioral experiment. Left: A rotating camera was used to rotate a stack 360° to display the 3-dimensional appearance of the configuration. Middle: Gravity's direction was randomly sampled from a spherical surface. Right: The physics simulator simulated the collapse trajectory of the stack under this selected direction, and participants reported whether the collapse trajectory was normal. b) The spherical surface of gravity's direction was determined by two parameters θ and φ. c) The procedure of calculating the normality ratio as the function of angle pairs. Left: Each cell represents the response of normality for an angle pair within a run. Middle: Responses for unsampled pairs were interpolated with the averaged responses along φ. Right: The normality ratio for each angle pair was calculated by averaging responses across runs and participants. d) Left: Gravity's direction encoded in the world model follows a Gaussian distribution with the vertical direction as the maximum likelihood. Note that the normality ratios for $\theta > 0$ were sampled from $\varphi \epsilon$ $(0°, 180°)$, and for $\theta < 0$ were sampled from $\varphi \epsilon$ $(180°, 360°)$. Right: The sphere represents the space of gravity's direction, with two poles pointing upward and downward, respectively. Each dot in the sphere represents one angle pair, and the color on a dot indicates the likelihood that the collapse trajectory under this gravity direction was judged normal. e) In a new setting, gravity's direction is reversed. Left: An example collapse trajectory when gravity's

direction was upward. Right: A trajectory when the direction was away from the vertical upward. f) Gravity's direction encoded in the world model when gravity's direction in the physical world was reversed. Error bar: standard error.

Contrary to this intuition, the subjective sense of the abnormality was not immediately apparent as θ moved away from 0; instead, the rate of reporting normality of collapse trajectories decreased gradually as a function of θ, which was the best fit by a Gaussian function with $\sigma = 19.9$ (Fig. 1d left). That is, the participants were 50.9% confident in reporting a normal collapse trajectory when the vertical offset of θ was 19.9°. In addition, accuracy in detecting the abnormality was not affected by φ (Extended Data Fig. 2), consistent with the uniformly distributed gravitational field in the physical world. This pattern was observed for all participants tested, with σ varying from 11.1 to 37.1 (Extended Data Fig. 2). Therefore, the world model on gravity is unlikely to be a faithful replica of the physical world; instead, it encodes gravity's direction as a Gaussian distribution with the vertical direction as the maximum likelihood (Fig 1d right).

To further test whether the world model on gravity, once established, is encapsulated from visual experience and task context, we inverted the virtual environment upside down with gravity's direction pointing upward, and then asked the same group of participants to judge whether collapse trajectories were normal (Fig 1e, see Supplementary Movie S2). We found that the normality ratio also decreased gradually as a function of θ (Fig. 1f, $\sigma = 17.2$; Extended Data Fig 3 for each participant), which was not significantly different from that in the environment with gravity pointing downward. Indeed, each participant's σ in the upright condition was in high agreement with the σ in the upside-down condition ($r = 0.91$, $p < 0.01$). That is, the visual experience and task context apparently did not cognitively penetrate humans' world model on gravity, suggesting that it is likely encapsulated as a cognitive module.

How does the stochastic gravity's direction in the world model affect our inference on objects' stability? To answer this question, we recruited an independent group of participants to estimate the stability of 60 stacks of different configurations (Fig 2a), half of which were stable. During the experiment, the participants were required to judge how stable each stack was on a 0-7 scale without feedback, which was used to index their subjective sense about stacks' stability. Two world models were constructed for comparison. One world model was equipped with a vertically downward direction of gravity without any stochastic variance. This deterministic model is intended to simulate how the stacks fell in the real world, and is therefore called a natural gravity simulator (NGS) (Fig 2b top). The other model is the same as the NGS, except that the deterministic direction of gravity in the NGS was replaced by the stochastic direction obtained from the previous psychophysical experiment. This model is thus called the mental gravity simulator (MGS, Fig 2c top). Both models were used to quantify the degree of stability by measuring the proportion of unmoved blocks after the collapse, where the proportion of unmoved blocks after the simulation was used to estimate the stability of the stacks.

NGS-estimated stability was significantly correlated with participants' subjective sense (Fig 2b bottom; $r = 0.70$, $p < 0.01$), consistent with previous findings (Battaglia et al., 2013). However, the participants were more inclined to judge stacks as more likely to collapse, as the dots in Fig 2b are more concentrated on the lower side of the diagonal line. This phenomenon is referred to as the inference bias, which was indexed as the difference in stability estimates between the participants and the NGS (inference bias = -0.31, $p < 0.01$) (see Methods). In other words, the participants were unlikely to infer stacks' stability from simulations with a deterministic direction of gravity pointing vertically downward. In contrast, the MGS randomly sampled pairs of (θ_s, φ_s) from the Gaussian distribution as gravity's directions 100 times, and the estimated stability of a stack was the averaged stability of simulations with different angle pairs. Aside from a similar magnitude of the correlation in the stability estimates between the participants and the MGS (Fig 2c bottom; $r = 0.75$, $p < 0.01$), the MGS, unlike the NGS, perfectly captured participants' judgment of stability because the points were evenly distributed along the diagonal line (inference bias = 0.04, $p > 0.05$; see Extended Data Fig. 4 for the agreement when the MGS was implemented with different Gaussian functions). In other words, the magnitude of the correlation coefficients is not the only indicator to evaluate the model's fitness. In short, the world model that represents gravity's direction as a Gaussian distribution around the vertical direction properly explains our tendency to judge stacks as more prone to collapse.

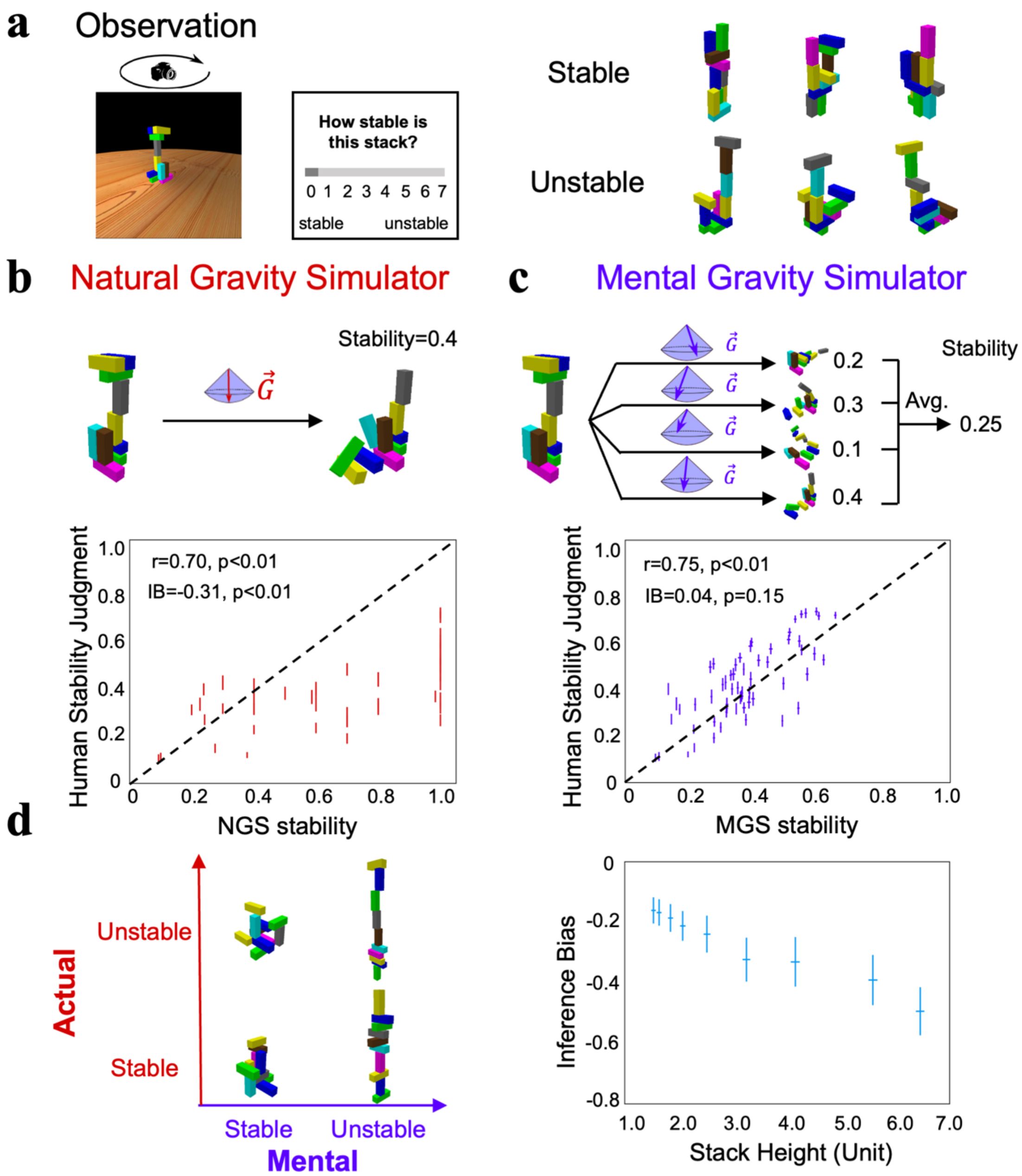


Fig 2. Stability inference by the world model on gravity. a) An experiment to rate the stability of stacks, half of which were stable and the other half unstable. b) Top: The procedure of the NGS to estimate the actual stability of stacks by simulation, and for unstable stacks the stability was indexed by the proportion of displaced blocks. Bottom: The correlation between the stability estimates of the participant and those of the NGS. Each dot represents one stack, and the lines denote the standard errors. c) Top: The procedure of the MGS, where the stability of a stack was estimated by averaging the estimated stabilities from multiple simulations with different gravity directions sampled from the Gaussian distribution. Bottom: The correlation between the stability estimates of the participant and those of the MGS. d) Left: The illusion that taller objects are perceived as more unstable than shorter ones. Right: The inference bias was indexed by the difference between the stability estimated by the MGS and that estimated by the NGS. The larger the negative values, the more likely stacks were

unstable. The x-axis denotes the height of a stack containing ten blocks, where the height, length, and width of each block were 1.2, 0.4, and 0.4, respectively. IB: inference bias. Error bar: standard error.

The stochastic world model illustrated by the MGS that led to participants' inference bias may explain the daily illusion that we perceive taller objects to be more unstable than shorter ones (Fig 2d left). An intuitive explanation from physics is that a tall object has a higher center of gravity, and thus an external perturbation makes it more likely to collapse. Our stochastic world model, on the other hand, provides an alternative explanation without introducing external perturbations, simply because deviations from gravity's veridical direction are likely to accumulate with the height of the objects. To test this conjecture, we constructed a set of stacks with different heights, and estimated the degree of stacks' stability with the MGS and the NGS, respectively. Because the MGS was considered to be the world model implemented in the brain, the inference bias here was calculated as the difference in stability estimates between the MGS and the NGS, with negative values indicating a tendency to judge a stable stack as an unstable one. Consistent with the inference bias found in humans, the MGS found stacks of all heights to be more prone to collapse (Fig 2d right; inference bias < 0, $p < 0.01$ for all heights). Critically, the bias increased monotonically with increasing height, consistent with the illusion that taller objects are considered more prone to collapse (see Extended Data Fig. 5 for the inference bias when the MGS was equipped with different levels of deviation). In short, the stochastic world model on gravity provides a more concise explanation for the daily illusion that taller objects are perceived as more likely to collapse, without assuming external perturbations.

The origin of the stochastic feature of the world model

A deterministic model that combines gravity's veridical direction with external perturbations, such as an external force or perceptual uncertainty (Allen et al., 2020; Battaglia et al., 2013; Lake et al., 2017; Smith & Vul, 2013), is theoretically equivalent to our stochastic model that represents gravity's direction in a Gaussian distribution; therefore, it also fits well with humans' inference on stability by fine-tuning the parameters of external perturbations. Although both the cognitive impenetrability and the self-consistency without resorting to an external perturbation found in our study favor the stochastic model over the deterministic one, more direct evidence comes from the origin of the stochastic feature of the world model.

Because our intelligence emerges and evolves under the constraints of the physical world, the stochastic feature may emerge as a biological agent interacts with the environment, where the mismatches between external feedback from the environment and internal expectations from the world model are in turn used to fine-tune the world model (Friston et al., 2021; MacKay, 1956; Matsuo et al., 2022). To simulate this process, here we designed a reinforcement learning (RL) framework to model this interactive process to illustrate how the world model on gravity evolves (Fig 3a). Specifically, an agent perceived a stack in the environment, which was then acted upon by a simulated gravity with direction parameters (i.e., θ and φ) sampled from a spherical direction space. The initial probabilities for the sampling directions were identical (Fig 3b, left). The final state of the stack served as the agent's expectation under the effect of the simulated gravity. The mismatch between the expectation and the observed final state of the stack under the natural gravity was used to update the sampling probability of the direction space, with a larger discrepancy leading to a larger decrease in probabilities through RL. Within this RL framework, we constructed 100,000 stacks of 2 to 15 blocks to train the world model on gravity. As the training progressed, the probabilities of the direction space gradually converged downward (Fig 3b, middle; see Extended Data Fig. 6 for the training trajectory). Although gravity's direction in the environment was vertical, the distribution of updated probabilities in the direction space was gradational (σ = 21.6; Fig 3b, right), which is close to gravity's direction represented in the world model derived from the psychophysics experiment on human participants. Therefore, the world model representing gravity's direction in a Gaussian distribution can emerge automatically as the agent interacts with the environment, without the need for any external perturbation.

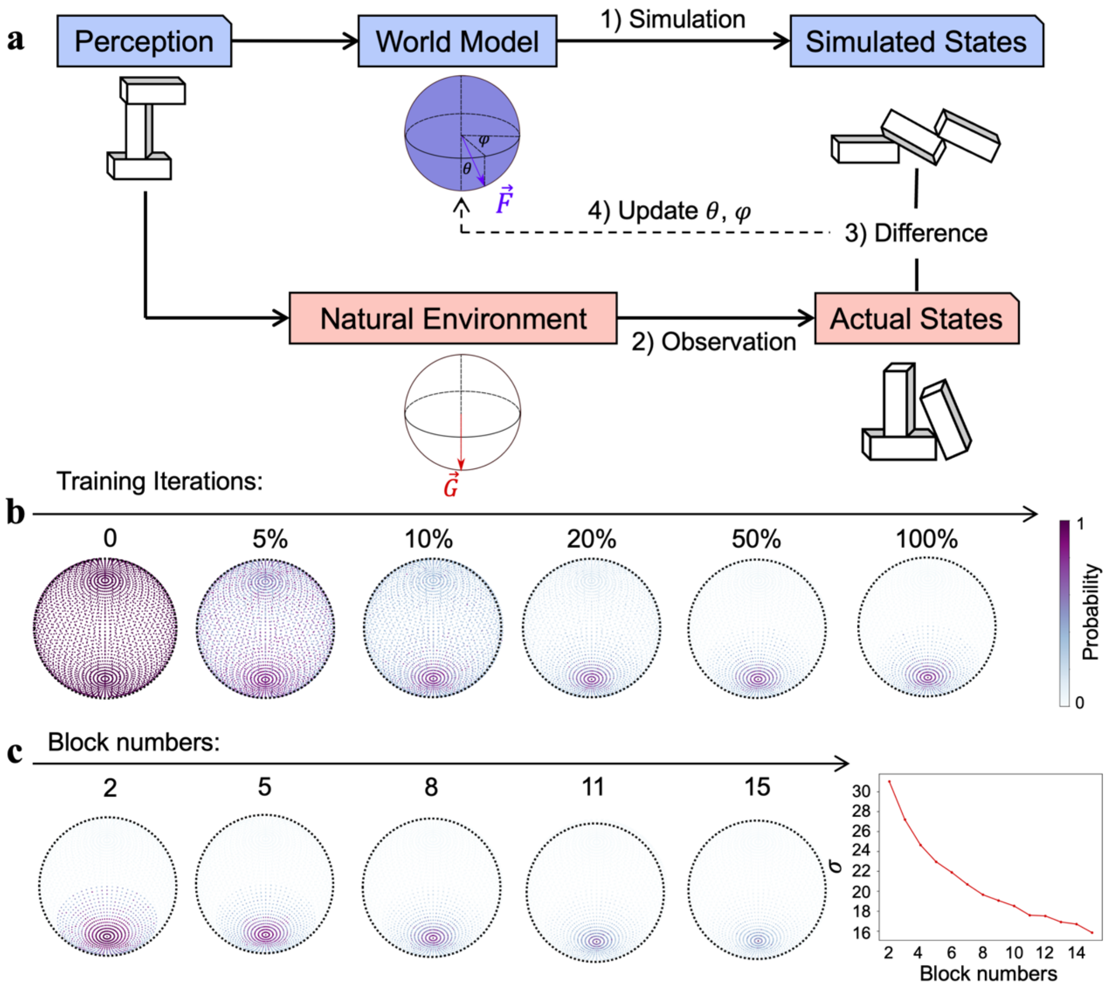


Fig 3. The origin of the stochastic feature of gravity's direction. a) The reinforcement learning framework, which updated gravity's direction (θ, φ) of the world model by minimizing the difference between the expectation from the internal simulation (i.e., simulated states) and the observation from the physical world (i.e., actual states). b) Gravity's directions, which were uniformly distributed on the spherical surface, gradually converged downward as the training progressed, and eventually stabilized in a Gaussian distribution with the vertical direction as the maximum likelihood. Color denotes the probability of a parameter pair being adopted as gravity's direction. c) Left: World models constructed by reinforcement learning when stacks in the physical world were composed of different numbers of blocks ranging from 2 to 15. Right: The variance of the Gaussian distribution, illustrated by the width of the distribution of gravity's direction on a spherical surface, monotonically decreased as the number of blocks in the stacks increased.

To further illustrate the idea that the environment constrains the form of intelligence, we systematically manipulated the appearance of the physical world while holding the natural gravity constant. Specifically, we constructed 14 worlds, each containing stacks of the same number of blocks, but with different configurations. The number of blocks ranged from 2 to 15. We trained the world model on gravity under the same RL framework for each world, and found that all world models represented gravity's direction in a Gaussian distribution (Fig 3c left; see Extended Data Fig. 7 for all world models). However, the width of the distribution, indexed by the parameter of σ, decreased monotonically as the number of blocks increased (Fig 3c right). This phenomenon was shown because in general stacks containing more blocks were more likely to be affected by forces whose directions were not perpendicular to the ground surface, which provided more information about gravity, and thus resulted in a more accurate representation of gravity's direction in the world model. In short, the world model on gravity resonates with not only the physical law governing the environment, but also the specific regularities of the environment the agent encountered.

The ecological advantage of the stochastic world model

When passing a cliff face, we have to be constantly aware of the stability of the rocks on the cliff. The ideal response would be both accurate and fast, but accuracy and speed are often difficult to achieve simultaneously. Here we investigated how the world model on gravity balances these two factors with its stochastic feature. To answer this question, we used a linear classifier (i.e., logistic regression) to model humans' decision-making behavior at different stages of the mental simulation. Specifically, we collected all the position coordinates of a stack's blocks at different stages of the simulation. The position difference between the intermediate states of the stack and the initial state provides information about the stability of the stack. For example, a stable stack should have no difference in the positions of the component blocks at all simulation stages, and an unstable stack should have a gradually increasing position difference. If the linear classifier detected the difference in positions sufficient for the classification at any stage, it classified the stack as unstable, otherwise stable (Fig 4a). The classification accuracy gradually increased as the simulation progressed until it reached the asymptote.

As expected, for the NGS (i.e., the world model with the deterministic direction of gravity), the accuracy at the plateau was close to 100% (95.3% on average, Fig 4b top red box), significantly higher than that for the MGS (80.1% on average, Fig 4b top blue box) ($t = 19.59$, $p<0.001$), simply because of the stochastic feature of gravity's direction. However, the MGS reached the plateau of decision accuracy much faster than the NGS (response time, indexed by the ratio between the time to reach the plateau and the time to reach the final stage: 27.1% vs. 75.2%, $t = 15.58$, $p < 0.001$) (Fig 4b middle). The same pattern was also observed with different variances of the Gaussian distribution (Extended Data Fig. 8). That is, the stochastic world model prioritized speed over accuracy, echoing the basic principle of survival: fleeing potential danger as quickly as possible, rather than making a perfect decision with a dreadful delay. In addition, by integrating the prediction accuracy and the response time as a measure of efficiency, we found that the stochastic world model provided a better balance between accuracy and speed, with an efficiency significantly higher than that provided by the NGS (3.49 vs. 1.32, $t = 9.12$, $p < 0.001$; Fig 4b bottom).

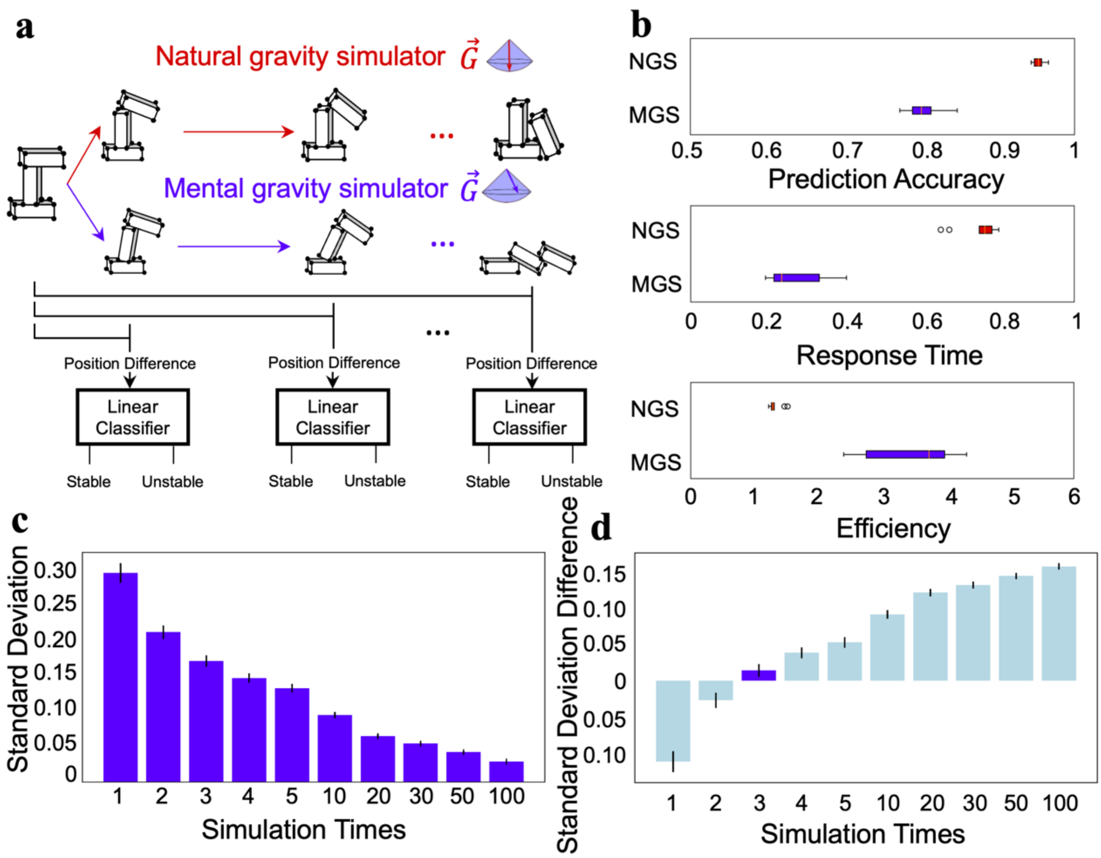


Fig 4. The ecological advantage of the stochastic feature. a) Illustration that modeled humans' decision-making behavior at different stages of the mental simulation using the NGS and MGS. b) The decision of the linear classifier based on the simulation of the MGS was less accurate than that of the NGS (top), but the decision was made faster in the MGS than in the NGS (middle). The MGS was more efficient than the NGS in combining accuracy and speed (bottom). c) The relationship between the number of simulations and the variance of the estimated stability. d) The difference in the variance

of the estimated stability between the participants and the MGS. The difference was minimal when the MGS ran the simulation three times. Error bar: standard error.

On the other hand, if time permits, multiple simulations with the MGS can significantly reduce the variance introduced by the stochastic representation of gravity's direction (Fig 4c). To explore whether humans adopted this strategy of performing multiple simulations before making a decision, we ran simulations with the MGS at different numbers of times and then matched them with humans' performance. We found that the variance of humans' inference on stability best matched that of the MGS after three simulations (Fig 4d; see Extended Data Fig. 9 for the model-behavior correspondence under different numbers of simulations). Therefore, humans are likely to run simulations a limited number of times to infer stacks' stability.

Discussion

In this study, we investigated how the physical law of gravity is embodied in the brain as a world model that guides inferences on objects' stability. A series of psychophysics experiments showed that the world model on gravity is not a faithful replica of the physical world, but rather a stochastic model that captures the essence of the vertically downward direction of gravity as the maximum likelihood of a Gaussian distribution. The stochastic feature of the world model not only fits humans' stability inference behavior better than the deterministic model, but also provides new insight into the daily illusion that taller objects are perceived as more likely to collapse. We further illustrated how the stochastic feature evolved through interactions with the environment using reinforcement learning, and well-balanced accuracy and speed to produce a unique ecological advantage for our survival in the physical world.

About 300 years ago, the philosopher Immanuel Kant proposed the intuition of space and time as *a priori* knowledge in the mind for us to understand the physical world (Kant, 1781), but only until recently have researchers investigated how the intuition is implemented in the brain as intuitive physics (Kubricht et al., 2017; McCloskey, 1983). In the Noisy Newtonian Framework, intuitive physics is depicted as a combination of Newtonian physics and uncertainty generated by noise (Battaglia et al., 2013; Kubricht et al., 2017; Sanborn et al., 2013). The introduction of uncertainty helps to reconcile the misconception occurring under unfavorable conditions, such as unfamiliar events or static scenes (Kaiser et al., 1986, 1992; Kim & Spelke, 1999; McCloskey, 1983; Smith & Vul, 2013), which was once thought to support Aristotelian physics (DiSessa, 1982; Halloun & Hestenes, 1985). The noise in previous studies was thought to originate from sources such as perceptual uncertainty or external perturbations of forces, rather than from the intuitive physical engine itself, which is thought to be a deterministic system. Our study extends these deterministic models by showing a stochastic world model that the noise instead came from the representation of gravity's direction under Gaussian distribution. The inherent stochastic feature of gravity's direction did not need to rely on external noise to explain the illusory instability of taller objects. In addition, it was also confirmed by the cognitive impenetrability of the Gaussian distribution of gravity's direction when gravity's direction in the physical world was reversed (Pylyshyn, 1980).

With a reinforcement learning framework, we further demonstrated a possible origin of the stochastic feature of the world model through interactions with the physical world. In contrast to summarizing statistical patterns from the experience (Bear et al., 2021; Li et al., 2016; Zhang et al., 2016), this framework was designed to simulate how an agent constructed the world model on gravity through agent-environment interactions. Specifically, a world model with undifferentiated directions of gravity generated a prediction on the stability of an object, and the mismatches between the prediction and the observation of the object from the physical world were used to fine-tune the distribution of the directions in the world model. This process is similar to how humans update their internal knowledge by comparing simulated expectations (Hegarty, 2004; Ullman et al., 2017) with actual observations (Baillargeon, 1994, 2004; Kotovsky & Baillargeon, 2000). After several generations of error minimization, a Gaussian distribution of gravity's direction with the vertically downward direction as the maximum likelihood was similar to that observed in the human world model. Interestingly, when the physical worlds that the agent interacted with changed their appearance with stacks of different heights, the world models maintained their general patterns, but the stochastic representation of gravity's direction changed accordingly. This finding not only demonstrates the robustness of the active inference (Hegarty, 2004; Ullman et al., 2017), which efficiently encodes critical features under different physical worlds, but also resonates with the idea that intelligence develops under the constraints of the physical world. Taken together, the finding from the RL framework implies that the world model on gravity in humans may also be constructed in the same way, possibly through the mechanism of the predictive coding in a generative process (Friston, 2018; Huang & Rao, 2011).

Our world model on gravity provides an example of the world model theory that emphasizes the predictive nature of generative neural networks implemented with *a priori* knowledge of the physical world (Friston et al., 2021; Land, 2014; Matsuo et al., 2022). In contrast to traditional discriminative neural networks that learn statistical patterns for stability from gigantic amounts of labeled stacks, generative models equipped with the physics laws governing the physical world rely much less on experience. Importantly, the stochastic feature of the model further enhances the efficiency by balancing accuracy and speed, which improves our chances of better survival (Cosmides & Tooby, 1997) and adaptation to novel environments (e.g., astronauts in outer space (Wang et al., 2022)). Indeed, the close link between human cognition and the physical world through interaction may shed light on the development of a new generation of AI with human-like intelligence that can work flexibly in open-ended environments (Marcus, 2018, 2020).

Methods

Creating stacks with different configurations

We designed a block-stacking procedure in a physical simulation platform (PyBullet) to generate stacks with different configurations. All stacks used in this study were generated using this procedure with the same parameters listed below.

The block-stacking procedure includes three steps (Extended Data Fig. 1a): (1) defining the designated area, (2) stacking blocks, and (3) fine-tuning block positions. The first step is to designate a restricted place area. All blocks of a stack were required to place within the designated area. The designated area controls the aggregation level of blocks, with a small area clustering blocks closer than a large area. The designated area is determined by two horizontal parameters x and y, which separately represent the size of the area in two horizontal directions. Therefore, when the block number is fixed, a smaller area in general constructs a higher stack. After designating the area, in step two we stacked blocks in random horizontal positions within the area one by one. If no block was positioned under a new block, the new block would be directly placed on the ground; otherwise, it would stack on the positioned block. The horizontal position of each block was independently sampled from a uniform distribution, with lower and upper bounds being $-x$ and $+x$, or $-y$ and $+y$ separately (x and y were all independently sampled from a uniform distribution $U(0.2, 2.0)$). The first two steps allow us to generate a large number of configurations within the designated area, which is the only restriction of the block-stacking procedure. To better control the physical stability of each stack, in step three we fine-tuned blocks in the stack by adjusting overlaps between every neighboring one, which was randomly sampled from a uniform distribution $U(0.2, 0.8)$. Smaller overlap between neighboring blocks is more likely to construct unstable stacks, whereas more extensive overlap results in more stable stacks. The overlap of neighboring blocks without contact is set to 0. Note that the overlap between neighboring blocks is not the only factor determining a stack's stability, and step three is used to generate stacks without consuming too many computational resources.

The size of each block has a 3D aspect ratio of 3:1:1 (length: width: height), with an arbitrary unit of 1.2:0.4:0.4. This constitutes three types of blocks (length, width, or height is 1.2, respectively, see Extended Data Fig. 1b). Each block of a stack was randomly selected as one of the three types of blocks. The mass of each block is set to 0.2 kg, and the friction coefficients and the coefficients of restitution between blocks are set to 1 and 0, respectively.

Estimating the stability of a stack

The stability of a stack was obtained by a rigid-body forward simulation under the natural gravity environment (i.e., natural gravity simulator, NGS). The direction of the natural gravity points downward (i.e., $\vec{G}$ = (0, 0, -9.8)), and all blocks of a stack are affected by the same gravity. Gravity is the only factor for changing the state of each block, and no external force is added during the simulation. Within each simulation, we recorded 500 simulation stages. In each stage, the center position of each block was collected to measure the stability of the stack. If the position of any block does not change during the simulation, the stack is considered stable, otherwise unstable. We formulate the stack's state according to the below criteria:

$$\begin{gathered} Stable\colon \forall t \wedge \forall m, |P_{tm} - P_{0m}| < \varepsilon \\ Unstable\colon \exists t \vee \exists m, |P_{tm} - P_{0m}| > \varepsilon \end{gathered} \tag{1}$$

Where t is a simulation stage, m is the block number of a stack, P_{tm} is the position of the block m at stage t, and ε is the just noticeable difference (i.e., j.n.d) of the perception, which is set to 0.01.

The stability of a stack is further calculated by measuring the proportion of displaced blocks, which is formulated as the following,

$$Stability = \frac{\sum_{m=1}^{M} \mathbb{I}(|P_{Tm} - P_{0m}| < \varepsilon)}{M} \tag{2}$$

Where M is the total number of blocks of a stack, and T is the final stage of the simulation (i.e., T = 500). $\mathbb{I}(\cdot) = 1$ when $|P_{Tb} - P_{0b}| < \varepsilon$, which denotes that the stack is stable.

Measuring participants' sensitivity to gravity's direction

We decomposed gravity's direction into three independent components (Fig. 1b).

$$\begin{gathered} G_x = g \sin\theta \cos\varphi \\ G_y = g \sin\theta \sin\varphi \\ G_z = g \cos\theta \end{gathered} \tag{3}$$

Where g is the magnitude of gravity (g = 9.8), which was fixed in this study. θ represents the vertical component, φ represents the horizontal component, and x, y, and z are three mutually perpendicular axes. The direction of the gravity was determined by the angle pair (θ, φ), where θ affects the extent of the collapse, and φ affects the orientation of the collapse. When θ is 0, gravity's direction is vertical.

We performed a psychophysics experiment to measure humans' sensitivity to gravity's direction. In this experiment, 10 participants (5 female, age range: 21-28) from Tsinghua University were recruited to finish four runs of the behavioral experiment, which measured their ability to detect the abnormality of stacks' collapse trajectories. The experiment was approved by the Institutional Review Board of Tsinghua University, and informed consent was obtained from all participants before the experiment.

The collapse trajectory of a stack was solely determined by gravity with different directions, where larger values of θ and φ made the trajectories more abnormal. A pilot experiment showed that almost all θ_s greater than 45 degrees made the collapse trajectory abnormal to most participants, and therefore in the experiment, θ ranges from 0 to 45 degrees with a step of 3 degrees. φ ranges from 0 to 360 degrees with a step of 24 degrees. Therefore, θ and φ consists of 16 values, respectively, which were randomly combined into 96 pairs of (θ, φ) with each value repeating 6 times in each run. In a trial, an unstable stack was constructed, and then the camera rotated one circle to show the 3D configuration of the stack to participants (Supplementary Movie S1). The configuration was randomly selected from a dataset with more than 2,000 unstable stacks, which was generated with the block-stacking procedure before the experiment. Each stack in the database was constructed with 10 blocks, and the color of each block was randomly rendered. There was a 1-sec delay after the rotation, during which the participants were instructed to infer the collapse trajectory based on the configuration. Then, simulated gravity with a direction determined by an angle pair (θ, φ) was applied to the stack, and the stack started to collapse. If the collapse trajectory met participants' expectations, they were instructed to choose 'Normal,' otherwise 'Abnormal'. Once the judgment was made, the subsequent trial started immediately. Each trial lasts about 10 seconds, taking 16 minutes for a run.

In addition, to test if participants' sensitivity to gravity's direction is encapsulated from visual experience and task context, we flipped gravity's direction upside down by inverting the camera's view, and the rest procedure remained the same.

To calculate participants' sensitivity to gravity's direction, we converted their behavioral judgment into normality ratio, which is the percentage that a trajectory was judged as normal, which was calculated as below:

$$Ratio_{\theta,\varphi} = \frac{n_{\theta,\varphi}}{N_{\theta,\varphi}} \tag{4}$$

Where $n_{\theta,\varphi}$ is the number of trajectories that were judged as 'Normal' with the angle pair (θ, φ), $N_{\theta,\varphi}$ is the total number of trajectories with the same angle pair. Because the angle pairs tested were a subset of all possible angle pairs, we used the average ratio along φ as the ratio of angle pairs untested (Fig. 1c) to acquire each participant's tuning curve. Finally, we calculated participants' sensitivity by fitting their normality ratios at different θ to a Gaussian distribution.

$$Ratio_{\theta} = Ae^{-\frac{\theta^2}{2\sigma^2}} \tag{5}$$

Where $Ratio_{\theta}$ is the normality ratio of θ, which was calculated by averaging the normality ratio along all φ_s, A is the magnitude of the gaussian curve, σ is the variance of the Gaussian curve. The best-fitted σ was used to index participants' sensitivity to gravity's direction, and a larger σ indicates a lower sensitivity.

Measuring participants' ability on stability inference

Another group of 11 participants (5 female, age range: 21-32) from Tsinghua University completed a behavioral experiment for judging the stability of 60 stacks. The experiment was approved by the Institutional Review Board of Tsinghua University, and informed consent was obtained from all participants before the experiment. One male participant (age: 25) was excluded from further analyses because his judgment showed an extremely weak correlation with the actual stability of stacks ($r_s < 0.30$ for all experimental runs), as compared to the rest of the participants.

The stacks contained 26 unstable and 34 stable stacks, which were randomly interleaved in each run. The participants were instructed to judge stacks' stability on an 8-point Likert scale, with 0 referring to 'definitely unstable' and 7 to 'definitely stable.' There was no feedback after each judgment. The participants completed six runs, within which the same group of stacks was presented but the sequence, blocks' colors, and camera's perspective were all randomized. After the experiment, only two participants reported that they suspected a few stacks were repeated in different runs, but they could not locate the stacks they suspected. Besides, their behavioral performance was not significantly different from other participants.

Participants' stability judgment was rescaled to 0 and 1 to match the scale of the stacks' stability. The participants' inference bias (IB) was indexed as the difference in stability judgment between the participants and the NGS, shown as

$$IB = Stability_{human} - Stability_{NGS} \tag{6}$$

Negative IB indicates that participants tended to consider a stable stack as an unstable one.

Estimating the stability of stacks based on the stochastic world model on gravity

The actual stability of a stack can be calculated with a one-time simulation of NGS ($\vec{G}$ = (0, 0, -9.8)). In contrast, the stochastic nature of mental gravity requires a multiple-time simulation with different gravity's directions. Specifically, we first randomly sampled several angle pairs (θ_s, φ_s) from the Gaussian distribution of gravity's directions in humans. The distribution was the average of two distributions acquired from the real world (i.e., gravity's direction is downward) and the inverted world (the direction is upward), with angles having larger normality ratios more likely being sampled. We then applied the simulated gravity with these sampled directions to the stack, and used the averaged stability with these directions as the stability of the stack estimated by the MGS. Similar to the IB between the participants and the NGS, the IB between the MGS and NGS was calculated as

$$IB = Stability_{MGS} - Stability_{NGS} \tag{7}$$

Stacks of different heights were created to investigate whether the stochastic world model on gravity results in the illusion that tall objects are considered less stable than short ones. The height of a stack was correlated with the size of the designated area, with a smaller area size corresponding to taller stacks. Therefore, we designated several square areas with different sizes. The side length of the squares ranged from 0.2 to 2.0, with an increase of 0.1. For each square, we used the block-stacking procedure to generate 100 stable and 100 unstable stacks consisting of 10 blocks. The height of each stack was the height of the highest block.

Investigating the origin of the stochastic world model on gravity

A reinforcement learning (RL) framework was used to simulate the development of the stochastic nature of the world model on gravity. To do this, we first created stacks whose block number ranged from 2 to 15 with the block-stacking procedure, and initialized a spherical force space, where θ ranged from 0 to 180 degrees and φ from 0 to 360 degrees. The spherical space covered all possible force directions, with the initial probability of being sampled by the MGS identical. During the training, three angle pairs (θ_s, φ_s) were sampled according to the probability of the spherical space, and then applied to a stack for simulating its collapse trajectory, which was divided into 500 stages. We optimized the sampling probability of gravity's direction by comparing the estimated stability (i.e., expectation) with the actual stability (i.e., observation) as a Q value, with a higher Q value suggesting that the sampled gravity's direction more likely mismatched the actual gravity's direction. The Q value was calculated as

$$Q = \frac{\sum_{m=1}^{M} \mathbb{I}(\left|P_{m,(\theta,\varphi)} - P_m\right| < \varepsilon)}{M} \quad (8)$$

Where $P_{m,(\theta,\varphi)}$ is the final position of block m with gravity's direction (θ, φ), P_m is the final position of block m with NGS, M is the block number of the stack, and the j.n.d. ε is set to 0.01. The mismatch between the expectation and the observation was used to update the sampling probability of the angle pair using a temporal difference optimization

$$W_{\theta,\varphi} \leftarrow W_{\theta,\varphi} + \gamma(Q - W_{\theta,\varphi}) \quad (9)$$

Where $\gamma = 0.15$ as the learning rate. This process was iterated to update the sample probability of angle pairs (θ_s, φ_s) until the training stopped. We prepared 100,000 configurations for the training.

Evaluating the ecological advantage of the model

To investigate how the world model on gravity balances response accuracy and speed, we trained a linear classifier (i.e., logistic regression) to model humans' decision-making process at different simulation stages. During the simulation, the same stack was separately simulated using the NGS and MGS, and we collected the position coordinates of all blocks at each stage. Differences in the positions of the blocks between the intermediate stage and the initial stage provided information about the stability of a stack, with more displaced blocks suggesting the lower stack's stability. As the simulation proceeded, differences in position gradually accumulated for unstable stacks, otherwise unchanged for stable stacks. The linear classifier was trained to judge whether a stack is stable with differences in position as inputs.

We used the block-stacking procedure to create stacks consisting of 2 to 10 blocks, and estimated their stabilities with the NGS for simulation in 500 stages. For each block number, there were 100 stable and 100 unstable stacks to train the linear classifier, and its prediction accuracy was measured with another group of 100 stable and 100 unstable stacks at every simulation stage.

The difference in positions of each block between the intermediate and initial stages was used as the input of the linear classifier. Specifically, we collected all vertex positions of a block during the simulation to acquire the difference in position, which included 8 coordinate points for each block in each stage. We did not collect the central position as previously used in the stability estimation, simply because it did not provide information on the shape and size of the block. We separately performed the simulation using the MGS and NGS, calculated the difference in position between the intermediate stage and the initial stage, and then flattened the difference to generate 24 position features for each block (i.e., eight positions per block in three-dimensional space). Therefore, for a 10-block stack as an example, there were 240 position features were prepared as the input of the linear classifier.

Prediction accuracy at each stage was estimated by evaluating whether a stack tested was stable with the MGS or with the NGS. The highest accuracy in the whole simulation stages was used as the prediction accuracy. Accordingly, the first simulation stage to reach the maximum accuracy provided information on response speed: reaching the maximum accuracy with a smaller number of stages indicates the classifier model accomplishes stability inference in a shorter amount of time (i.e., quick response). Therefore, we measured the response speed by estimating the steps to reach the accuracy plateau.

$$Time = \frac{\hat{t}}{T}$$
$$\hat{t} = arg \max_{t} Accuracy_t \tag{10}$$

Where $Accuracy_t$ is the accuracy of stage t. $\hat{t}$ is the stage that a linear classifier acquires the maximum accuracy for the first time, T is the total stage number of each simulation (T = 500). Higher values indicate longer response time (i.e., slower response). Finally, the efficiency of the stability inference, which is the balance between accuracy and speed, by dividing the prediction accuracy by the response time.

$$Efficiency = \frac{Accuracy}{Time} \tag{11}$$

References

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

Baillargeon, R. (1994). How do infants learn about the physical world? *Current Directions in Psychological Science*, *3*(5), 133–140.

Baillargeon, R. (2004). Infants' physical world. *Current Directions in Psychological Science*, *13*(3), 89–94.

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

Bear, D. M., Wang, E., Mrowca, D., Binder, F. J., Tung, H.-Y. F., Pramod, R., Holdaway, C., Tao, S., Smith, K., Sun, F.-Y., & others. (2021). Physion: Evaluating physical prediction from vision in humans and machines. *ArXiv Preprint ArXiv:2106.08261*.

Blatner, D. (2013). *Spectrums: Our mind-boggling universe from infinitesimal to infinity*. A&C Black.

Cosmides, L., & Tooby, J. (1997). *Evolutionary psychology: A primer*.

Coumans, E., & Bai, Y. (2016). *Pybullet, a python module for physics simulation for games, robotics and machine learning*. https://pybullet.org

DiSessa, A. A. (1982). Unlearning Aristotelian physics: A study of knowledge-based learning. *Cognitive Science*, *6*(1), 37–75.

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

Friedman, W. J. (2002). Arrows of time in Infancy: The representation of temporal–causal invariances. *Cognitive Psychology*, *44*(3), 252–296.

Friston, K. (2018). Does predictive coding have a future? *Nature Neuroscience*, *21*(8), 1019–1021.

Friston, K., Moran, R. J., Nagai, Y., Taniguchi, T., Gomi, H., & Tenenbaum, J. (2021). World model learning and inference. *Neural Networks*, *144*, 573–590.

Halloun, I. A., & Hestenes, D. (1985). Common sense concepts about motion. *American Journal of Physics*, *53*(11), 1056–1065.

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

Huang, Y., & Rao, R. P. (2011). Predictive coding. *Wiley Interdisciplinary Reviews: Cognitive Science*, *2*(5), 580–593.

Indovina, I., Maffei, V., Bosco, G., Zago, M., Macaluso, E., & Lacquaniti, F. (2005). Representation of visual gravitational motion in the human vestibular cortex. *Science*, *308*(5720), 416–419.

Kaiser, M. K., Jonides, J., & Alexander, J. (1986). Intuitive reasoning about abstract and familiar physics problems. *Memory & Cognition*, *14*(4), 308–312.

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

Kant, I. (1781). *The Critique of Pure Reason*.

Kim, I.-K., & Spelke, E. S. (1999). Perception and understanding of effects of gravity and inertia on object motion. *Developmental Science*, *2*(3), 339–362.

Kotovsky, L., & Baillargeon, R. (2000). Reasoning about collisions involving inert objects in 7.5-month-old infants. *Developmental Science*, *3*(3), 344–359.

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

Lacquaniti, F., & Maioli, C. (1989). Adaptation to suppression of visual information during catching. *Journal of Neuroscience*, *9*(1), 149–159.

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

Land, M. F. (2014). Do we have an internal model of the outside world? *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1636), 20130045.

Li, W., Azimi, S., Leonardis, A., & Fritz, M. (2016). To fall or not to fall: A visual approach to physical stability prediction. *ArXiv Preprint ArXiv:1604.00066*.

MacKay, D. M. (1956). The epistemological problem for automata. In *Automata Studies.(AM-34), Volume 34* (pp. 235–252). Princeton University Press.

Marcus, G. (2018). *Deep Learning: A Critical Appraisal*.

Marcus, G. (2020). The next decade in ai: Four steps towards robust artificial intelligence. *ArXiv Preprint ArXiv:2002.06177*.

Matsuo, Y., LeCun, Y., Sahani, M., Precup, D., Silver, D., Sugiyama, M., Uchibe, E., & Morimoto, J. (2022). Deep learning, reinforcement learning, and world models. *Neural Networks*.

McCloskey, M. (1983). Intuitive physics. *Scientific American*, *248*(4), 122–131.

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

Pramod, R., Cohen, M. A., Tenenbaum, J. B., & Kanwisher, N. (2022). Invariant representation of physical stability in the human brain. *ELife*, *11*, e71736.

Pylyshyn, Z. W. (1980). Computation and cognition: Issues in the foundations of cognitive science. *Behavioral and Brain Sciences*, *3*(1), 111–132.

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

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

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

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.

Wang, Y., Zhang, X., Wang, C., Huang, W., Xu, Q., Liu, D., Zhou, W., Chen, S., & Jiang, Y. (2022). Modulation of biological motion perception in humans by gravity. *Nature Communications*, *13*(1), 1–10.

Zago, M., & Lacquaniti, F. (2005). Visual perception and interception of falling objects: A review of evidence for an internal model of gravity. *Journal of Neural Engineering*, *2*(3), S198.

Zago, M., McIntyre, J., Senot, P., & Lacquaniti, F. (2009). Visuo-motor coordination and internal models for object interception. *Experimental Brain Research*, *192*(4), 571–604.

Zhang, R., Wu, J., Zhang, C., Freeman, W. T., & Tenenbaum, J. B. (2016). A comparative evaluation of approximate probabilistic simulation and deep neural networks as accounts of human physical scene understanding. *ArXiv Preprint ArXiv:1605.01138*.

Zhou, L., Smith, K., Tenenbaum, J., & Gerstenberg, T. (2022). *Mental Jenga: A counterfactual simulation model of physical support*.

Acknowledgments

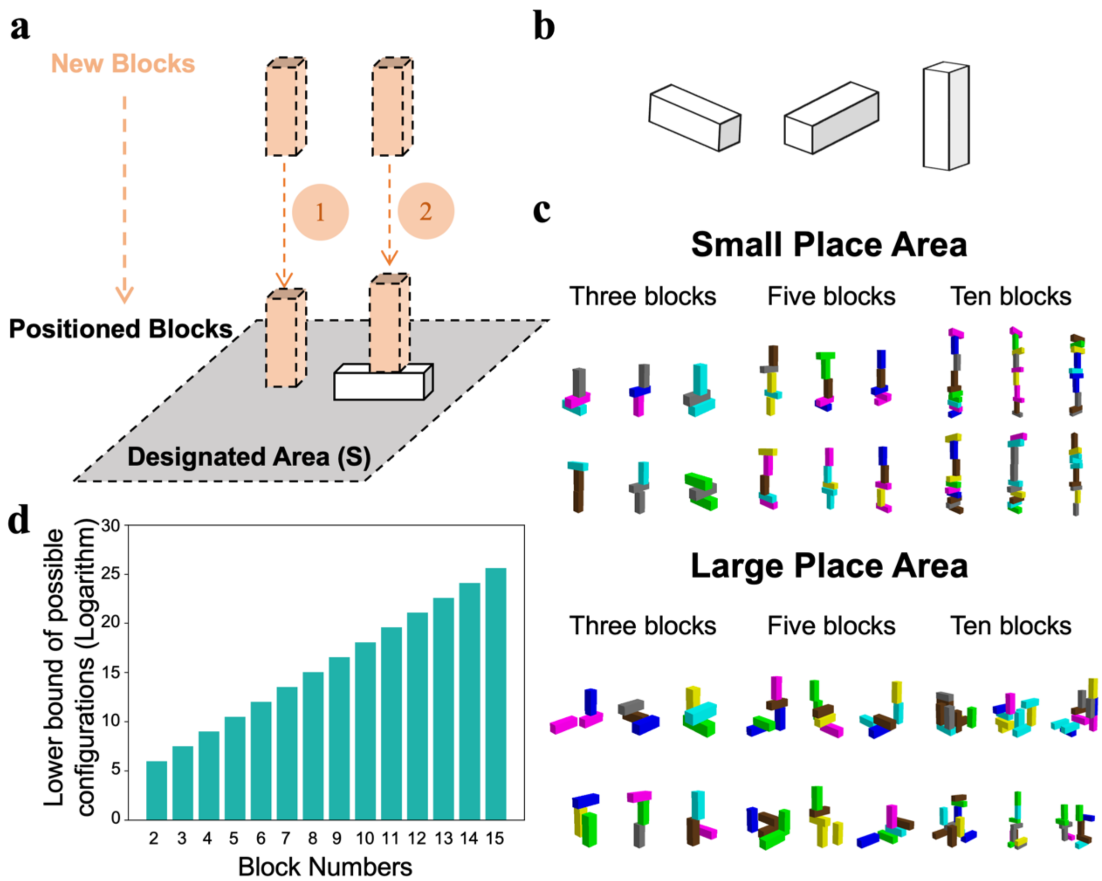
Funding: This study was funded by Beijing Municipal Science & Technology Commission and Administrative Commission of Zhongguancun Science Park (Z221100002722012), the Shuimu Tsinghua Scholar Program (T.H.), Tsinghua University Guoqiang Institute (2020GQG1016), Tsinghua University Qiyuan Laboratory, and Beijing Academy of Artificial Intelligence (BAAI).

Author contributions: J.L. conceptualized the study. T.H. designed and conducted the experiments. T.H. analyzed data. T.H. and J.L. wrote the manuscript.

Competing interests: Authors declare no competing interests.

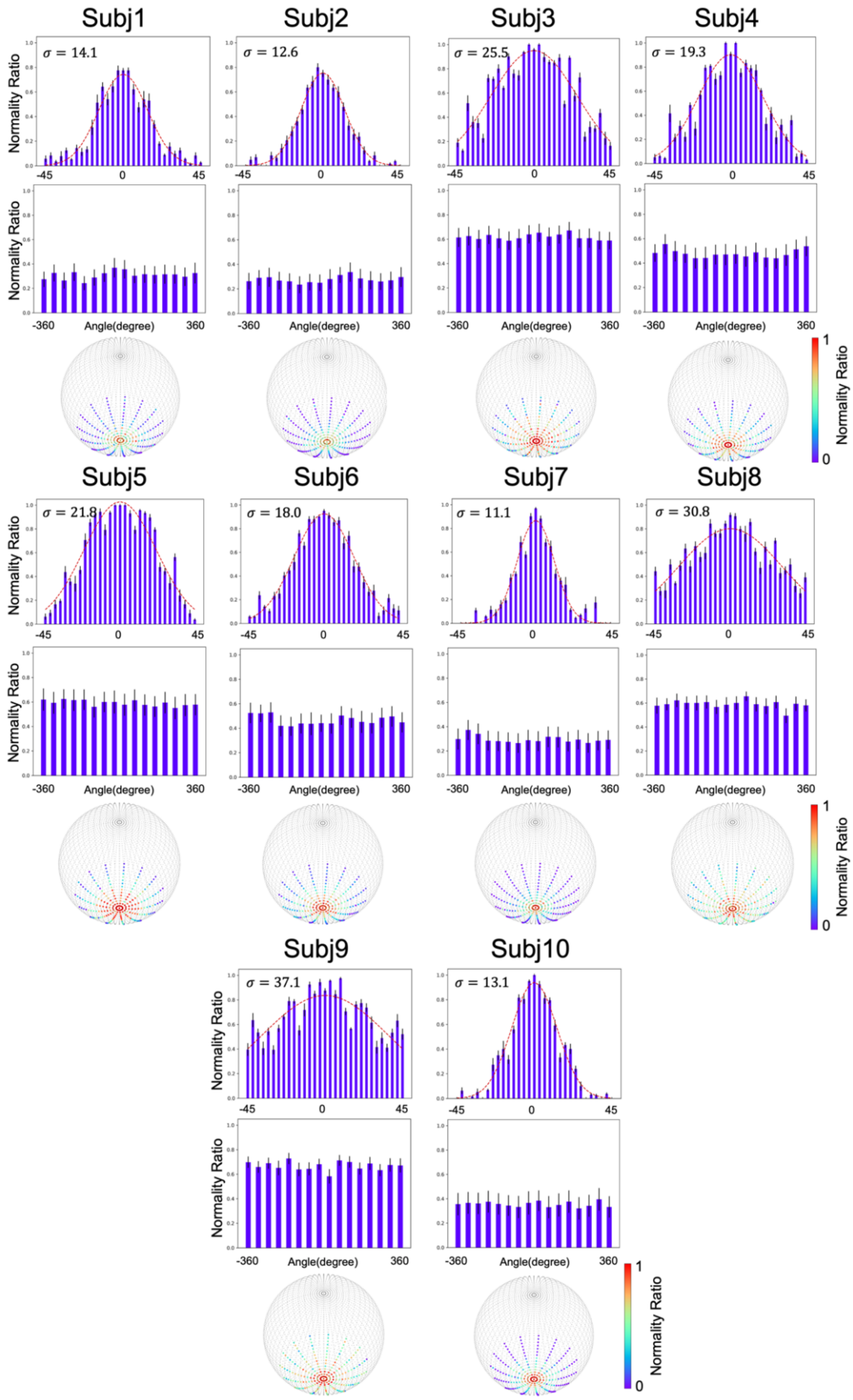
Data and materials availability: All code and data underlying our study and necessary to reproduce the results are available on Github: https://github.com/helloTC/GravityWorldModel.

Extended Data Fig.1



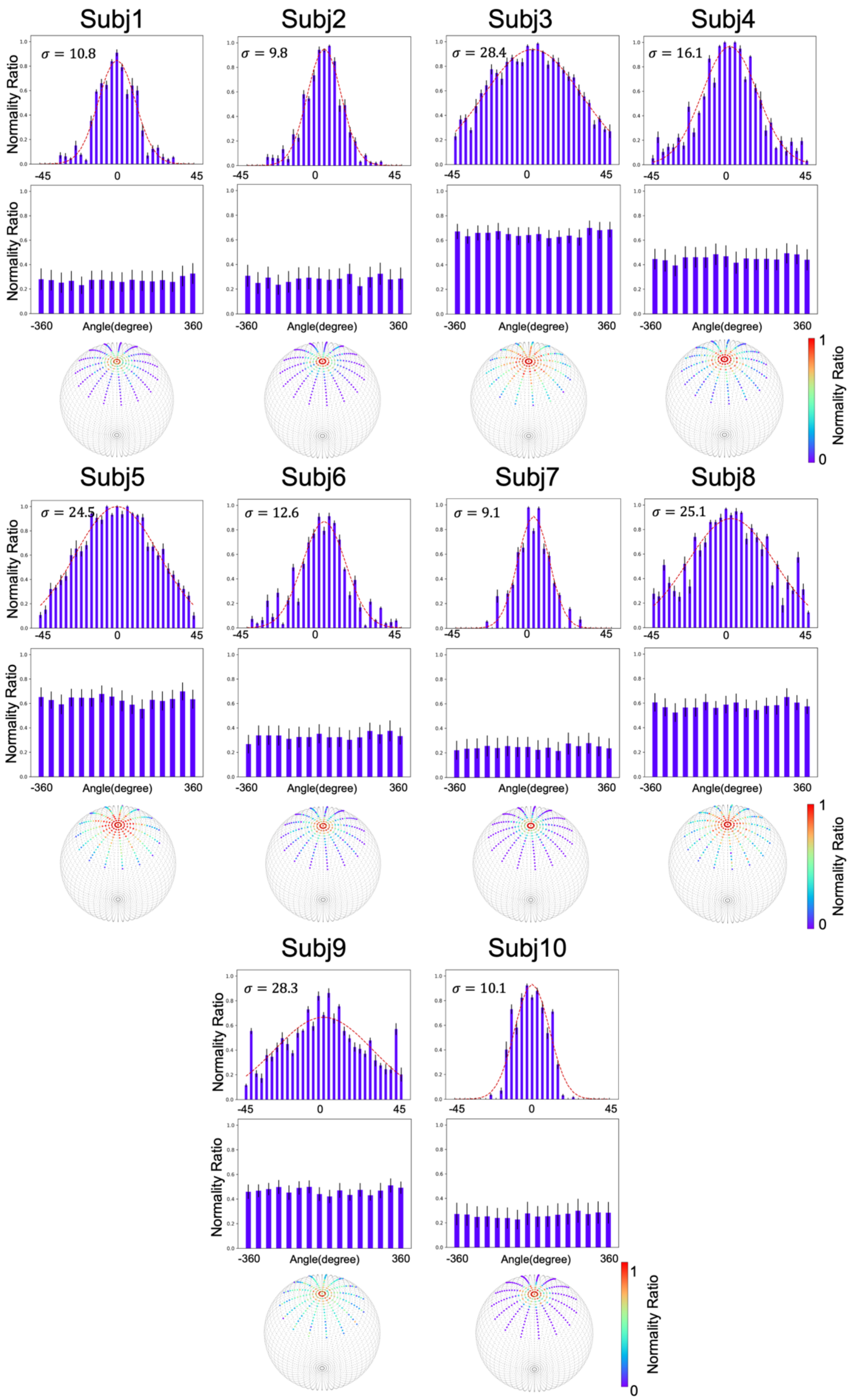
Extended Data Fig. 1 **Construction of stacks with different configurations.** a) Illustration of the block-stacking procedure to create stacks in different configurations. A configuration was constructed by placing multiple blocks within a designated area. If there was no positioned block in the area, a new block was placed on the ground; otherwise, it was placed on top of the positioned block. b) Three types of blocks with an aspect ratio of 3:1:1. c) This procedure can create a large number of stacks with different configurations within designated areas. Note that in small areas, the height of stacks was taller. d) The lower bound of configurations' possible number showed an exponential relation with the number of blocks in a stack. The procedure can create at least 3.72×10^{19} configurations for stacks consisting of 10 blocks. See the appendix for the estimation.

Extended Data Fig.2



Extended Data Fig. 2 **The stochastic world model on gravity of each participant.** The normality ratios of θ followed a Gaussian distribution, with the variance ranging from 11.1 to 37.1. No stochastic characteristic was observed in φ.

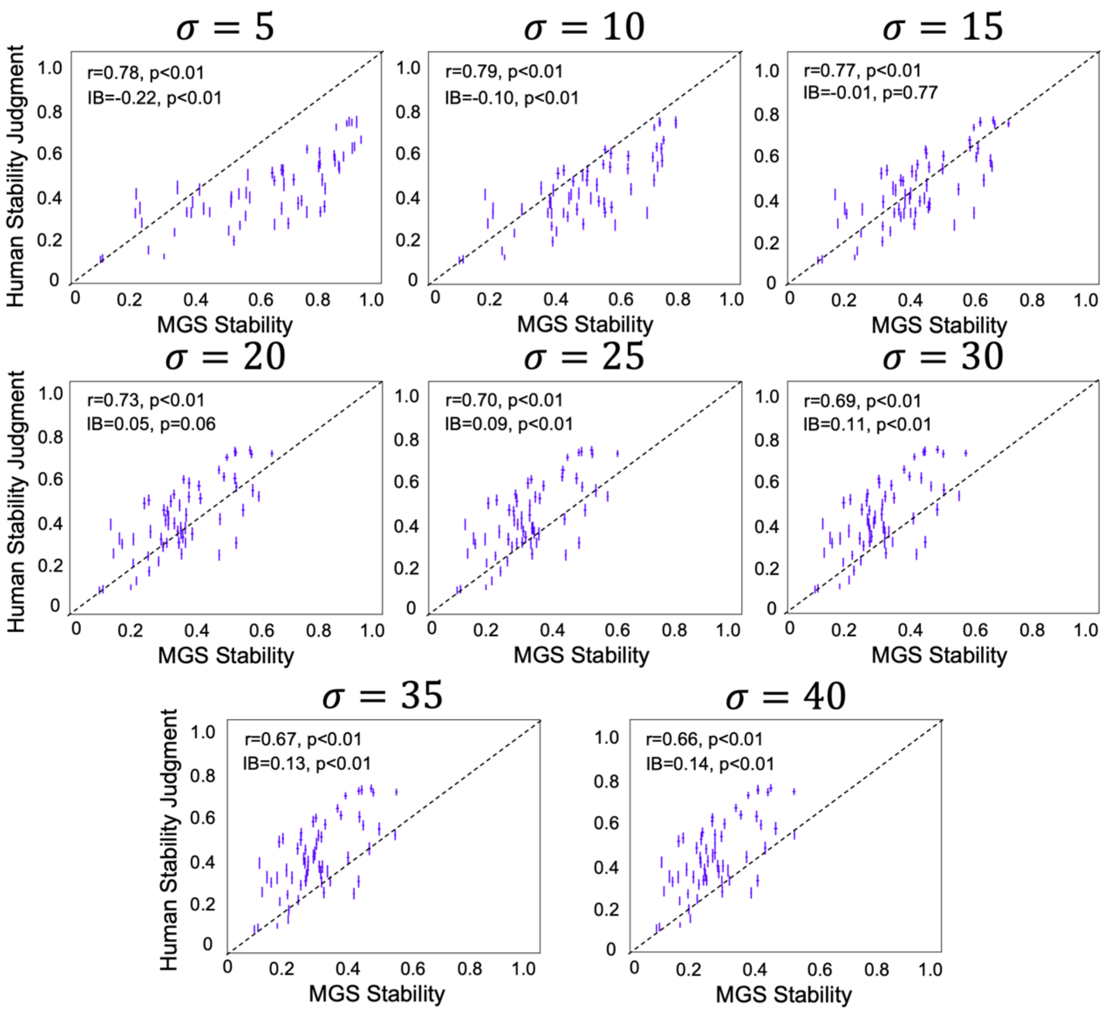
804 **Extended Data Fig. 3**



805

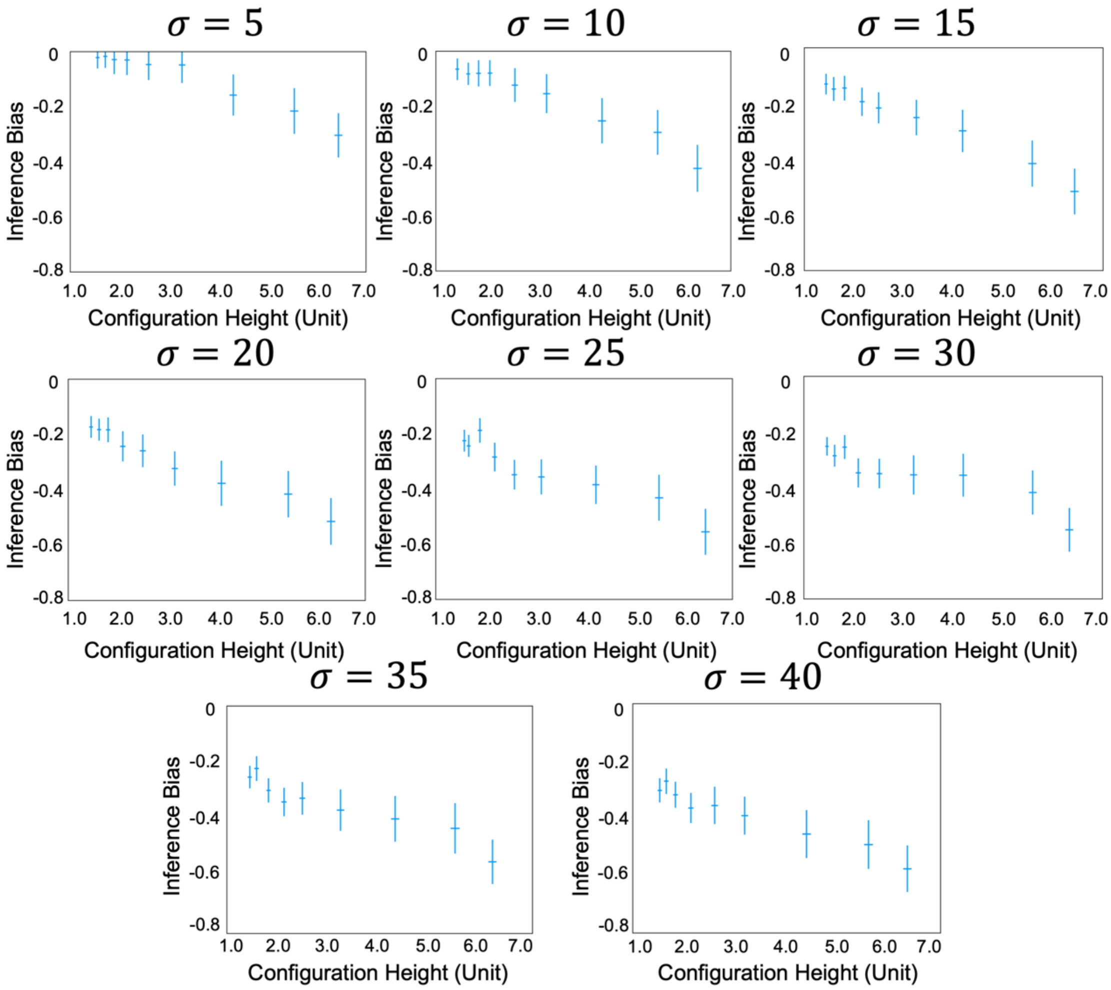
Extended Data Fig. 2 **The stochastic world model on gravity of each participant when gravity's direction was inverted.** The normality ratios of θ also followed a Gaussian distribution, with the variance ranging from 9.1 to 28.4, and no stochastic characteristic was observed along φ.

Extended Data Fig. 4



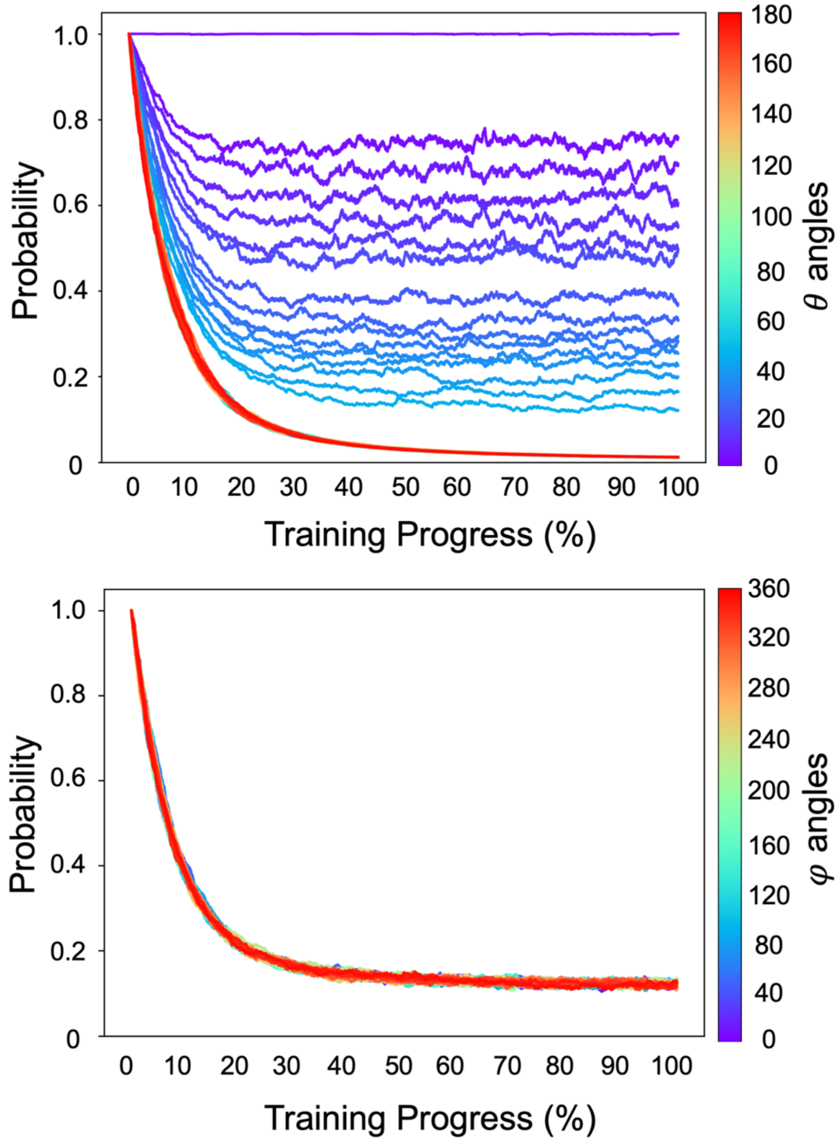
Extended Data Fig. 4 **Relation between the stability estimated by the MGS stability and that by participants when the world model was implemented with different Gaussian functions.** Only when the world model embodied Gaussian functions with intermediate variance (i.e., $\sigma \in (15,20)$) did the stability estimated by the MGS match participants' stability inference. On the other hand, when the variance was small, most points were positioned below the diagonal line, indicating the model considered stacks more stable in general as compared to participants' judgment. When the variance was large, the model considered stacks less stable. Note that all models showed high correlation coefficients regardless of the bias. In other words, the magnitude of the correlation is not the sole indicator to evaluate the fitness of the model. IB: inference bias.

Extended Data Fig. 5



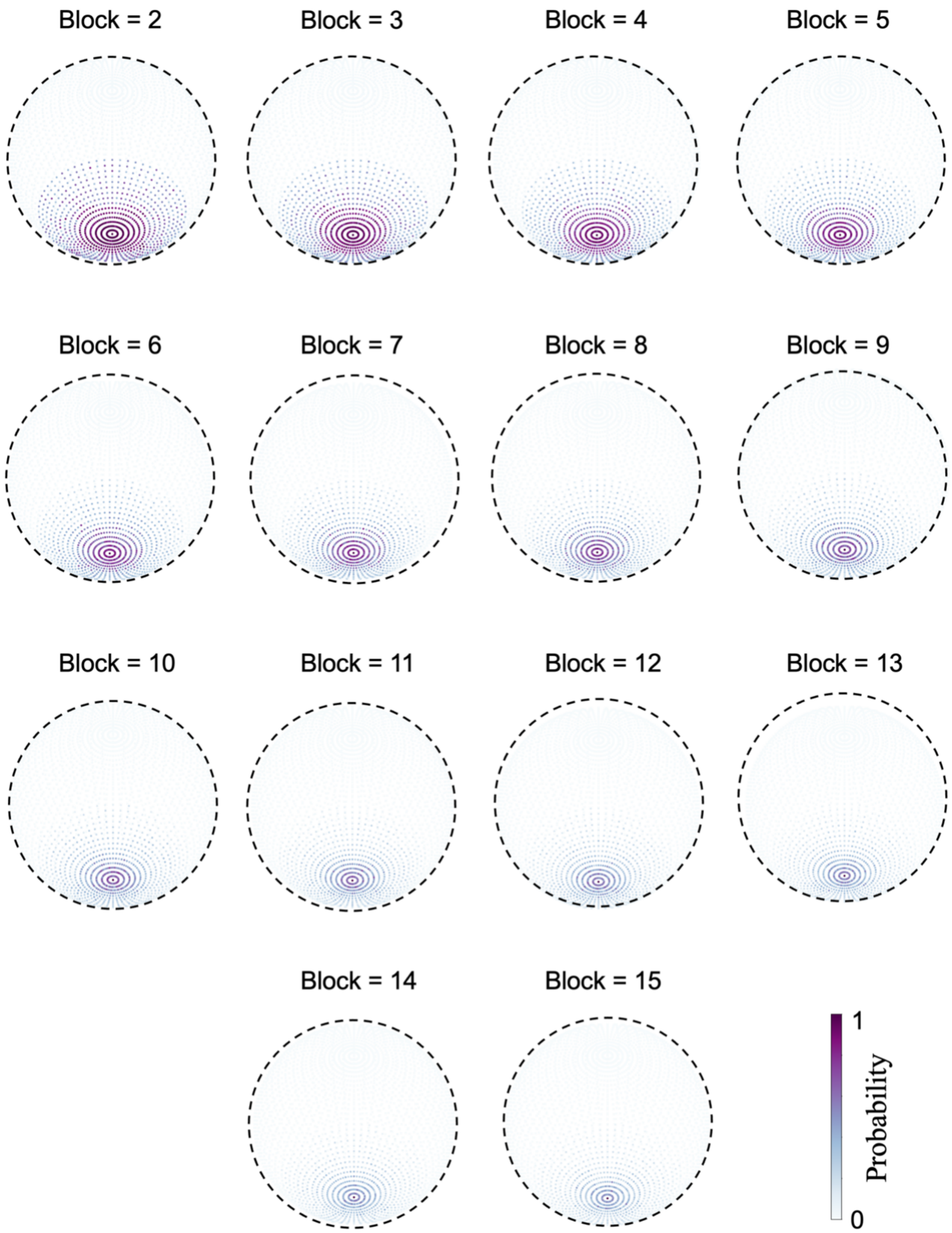
Extended Data Fig. 5 **Height illusion of stability inference when the world model was implemented with different Gaussian functions.** The illusion that tall objects are considered more unstable than short ones manifests at all levels of variances of Gaussian functions, with larger variance leading to a stronger illusion.

Extended Data Fig. 6



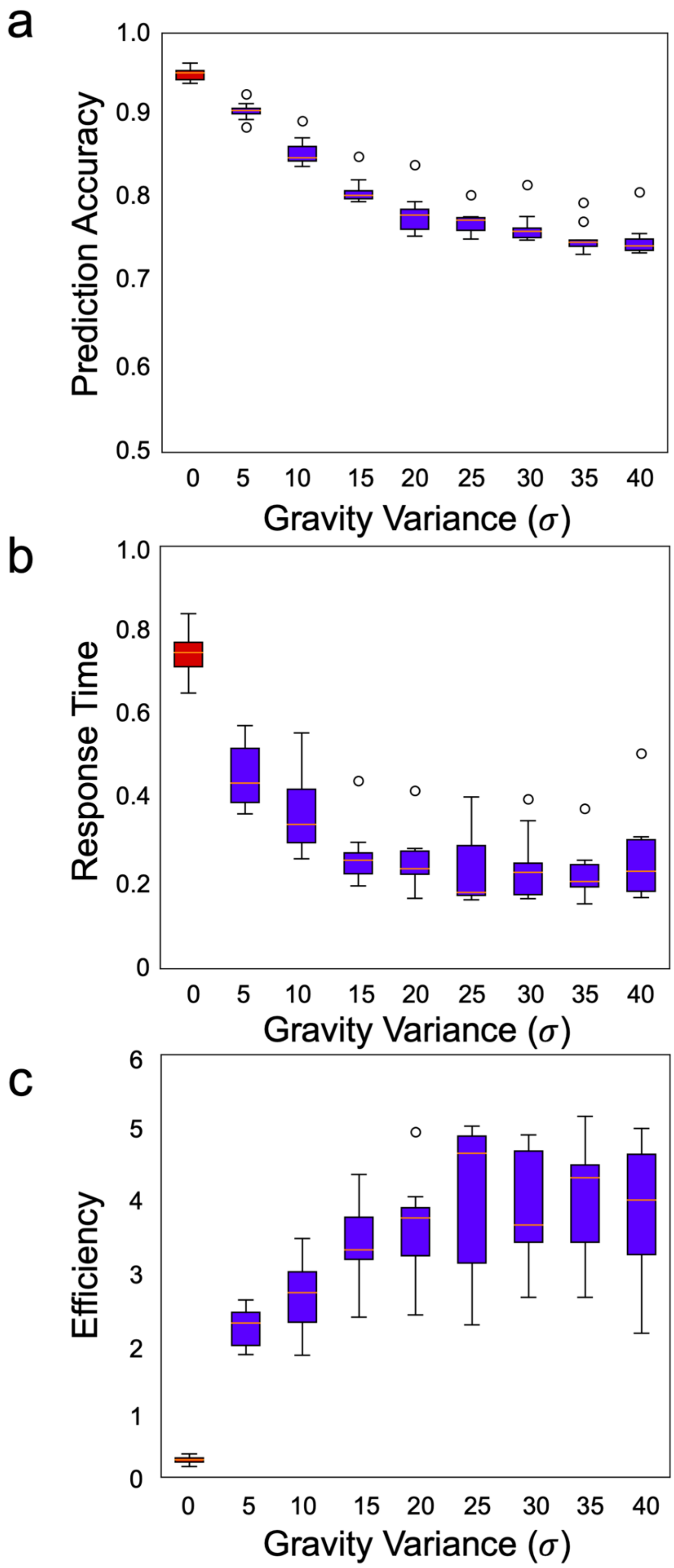
Extended Data Fig. 6 **The developmental trajectory of θ (Top) and φ (Bottom) angles.** Sampling probabilities of θ angles gradually decreased during reinforcement learning, with the probabilities from smaller θ angles having a lower decrement tendency. The probability of θ without any deviation (i.e., $\theta = 0$) keeps unchanged. Probabilities of all θ angles finally reached convergence after about 50% training progress. Different from θ angles, sampling probabilities of the φ angles dropped evenly.

Extended Data Fig. 7



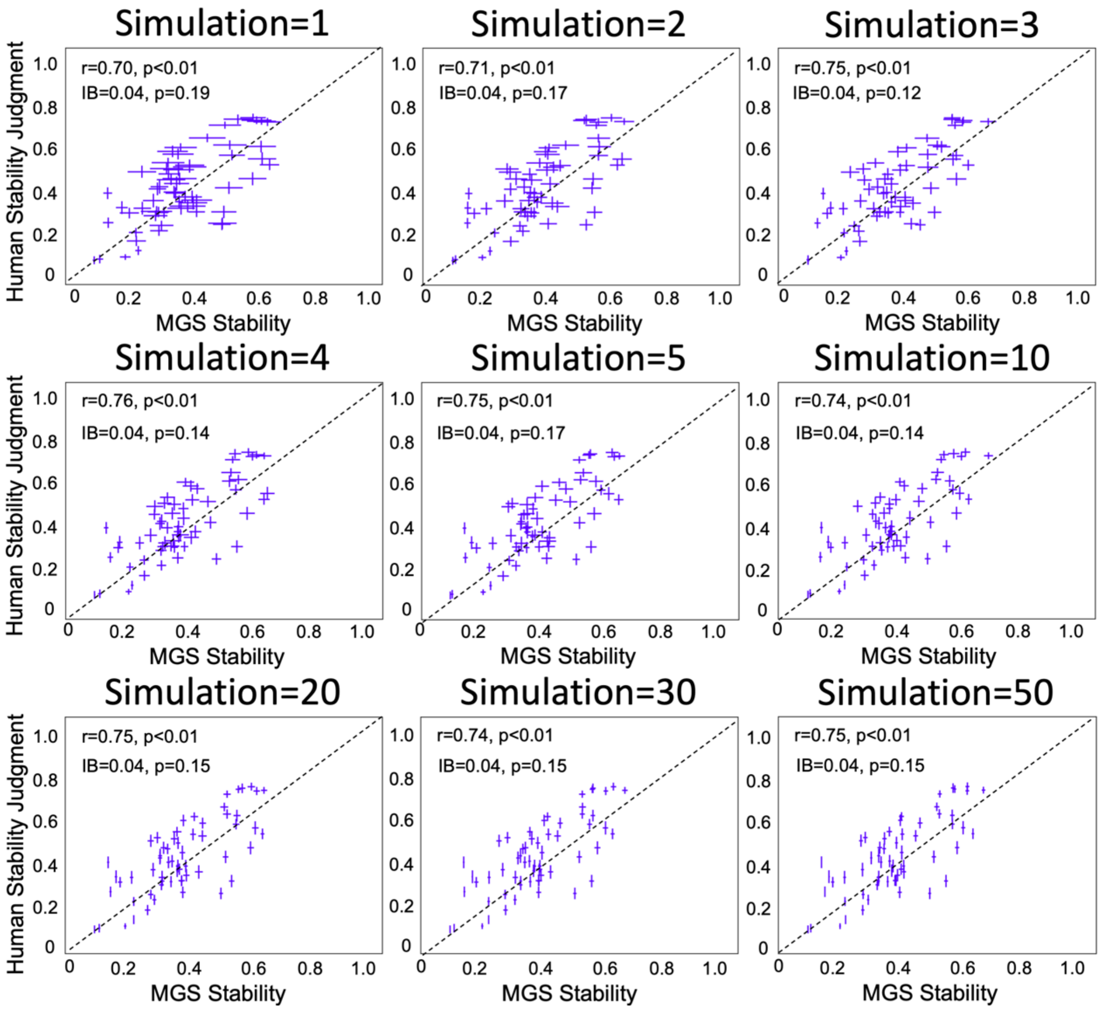
Extended Data Fig. 7 **The world models developed in the world containing stacks with different numbers of blocks.** The number of blocks ranged from 2 to 15, and in all the worlds gravity's direction was in Gaussian distributions with the vertical direction as the maximum likelihood. Note that the world with stacks consisting of more block numbers led to smaller variances in the Gaussian function.

Extended Data Fig. 8



Extended Data Fig. 8 **Ecological advantage of the world model embodied with different Gaussian functions.** a) Prediction accuracy decreased when the variance of the Gaussian function increased, and reaches an asymptote of 0.75. b) Response time decreased as the variance increased, and reached an asymptote of 0.20. c) The prediction accuracy and response time was combined as a measurement for efficiency, which gradually increased monotonically as the function of the variance until an asymptote of 4. Red box: the world model embodied no stochastic characteristic (i.e., the deterministic model); Blue box: the world model with different levels of variances. Error bar: standard error.

Extended Data Fig.9



Extended Data Fig. 9 **The relation between the number of simulations and the variance of stability inference.** The simulation showed that the variance of stability inference decreased with the number of simulations. Note that the variance in the world model observed in participants best matched the variance when the simulation of the MGS was conducted three times.

Appendix: Estimate the lower bound of the possible number of configurations

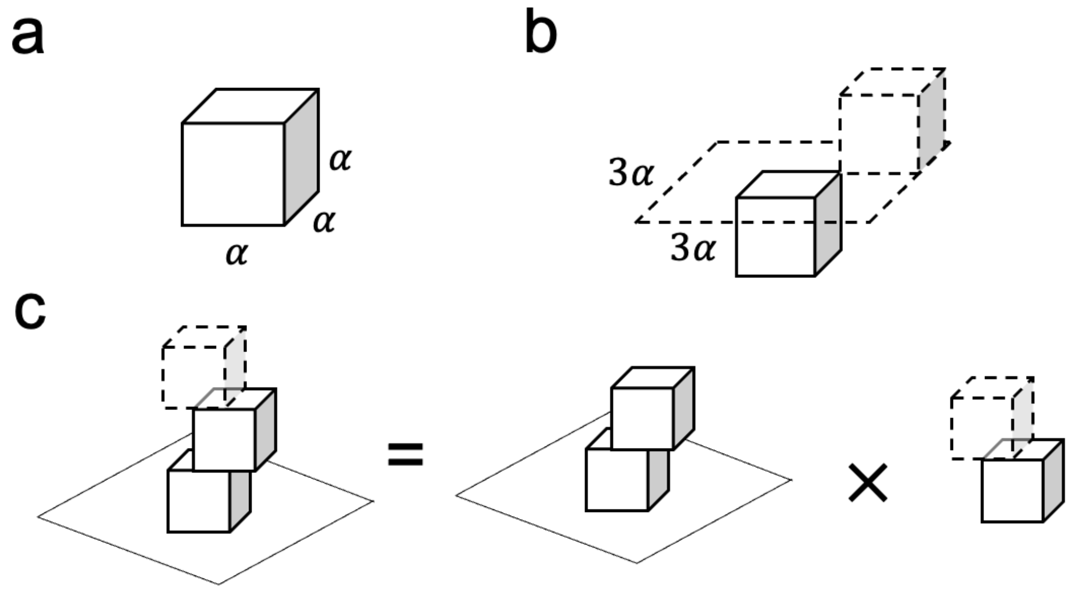
A configuration is a structure composed of several contact blocks. To simplify the computation of estimating the number of possible configurations, here we constrained the shape of blocks and the position where the blocks were placed.

The shape constraint*: the blocks used to form a configuration are all uniform rectangular blocks with the same aspect ratio.*

The position constraint*: only one block is allowed to be placed on the same layer of the configuration.*

Thus, the problem is then simplified to estimate the possible number of configurations when only one rectangular block with the aspect ratio of $\alpha:\beta:\gamma$ (i.e., **the shape constraint**) is allowed to place in one layer (i.e., **the position constraint**). Note that the constraints significantly reduce the number of estimated configurations.

We illustrated our solution by starting with a simple case: the aspect ratio of blocks is $\alpha:\alpha:\alpha$.



Appendix Fig 1. An illustration of the procedure to estimate the possible number of configurations when blocks have an aspect ratio of $\alpha:\alpha:\alpha$. (a) the cubic block with the length, width and height are α. (b) Constructing a configuration by stacking two cubic blocks. The upper block could only be placed within a $3\alpha \times 3\alpha$ area to guarantee contact with the lower block. (c) A three-block configuration can be viewed as stacking a cubic block on a two-block configuration.

The condition when the aspect ratio of blocks is $\alpha:\alpha:\alpha$

The block with the aspect ratio of $\alpha:\alpha:\alpha$ is a cube (Appendix Fig 1a). The side length of the cube is defined as α. Consider a configuration with two stacking blocks, the upper block needs to be placed in a $3\alpha \times 3\alpha$ area to ensure contact with the bottom block (Appendix Fig 1b). To estimate the possible number of this simple situation, we defined a visual acuity υ, which is the minimum resolution to distinguish two stacks (i.e., j.n.d.). Note that υ is a small value and here we set it as $\upsilon = 0.01$ to match the minimal position difference for stability estimation in the simulation platform (please see Methods). Therefore, the possible number of the configuration containing two cubic blocks is

$$N_{C2} = \left(\frac{2\alpha}{\upsilon}\right)^2 \tag{1}$$

Where N_{C2} indicates the possible number of configurations containing two cubic blocks.

We further consider the situation with more cubic blocks. For a stack that contains three cubic blocks, it can be viewed as placing a cubic block on a two-block stack (Appendix Fig 1c). Therefore, the total possible number of configurations is the multiplication of two two-block configurations, which is formulated as

$$N_{C3} = N_{C2} \times N_{C2} = {N_{C2}}^2$$

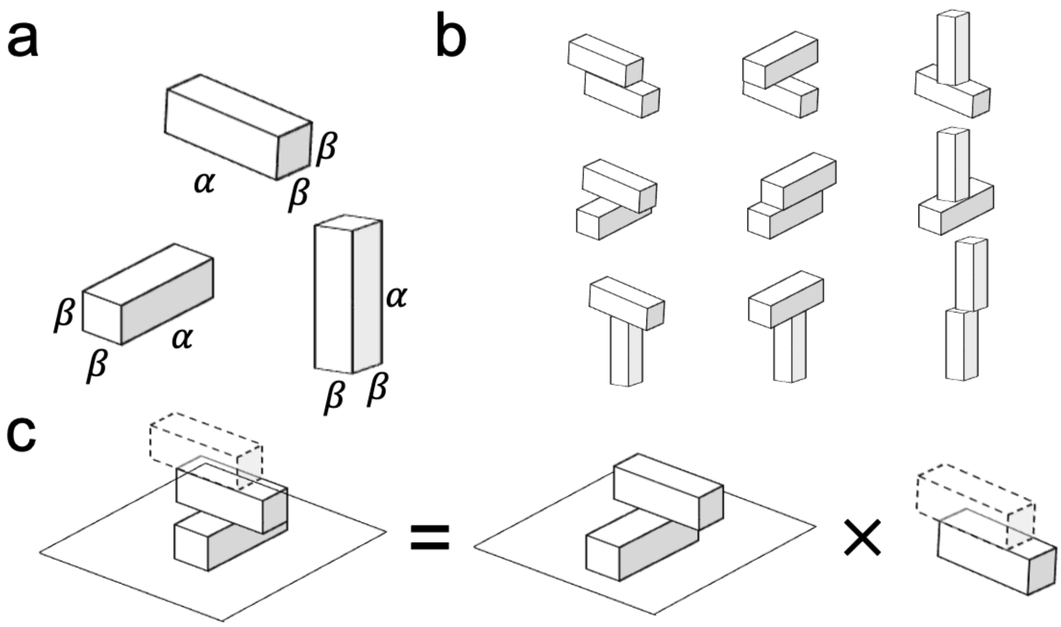
Similarly, the possible number of configurations for stacks containing four cubic blocks is

$$N_{C4} = N_{C3} \times N_{C2} = {N_{C2}}^3$$

Accordingly, the possible number of configurations with M cubic blocks is

$$N_{CM} = N_{C(M-1)} \times N_{C2} = \cdots = {N_{C2}}^{M-1} = \left(\frac{2\alpha}{\upsilon}\right)^{2M-2}, M \geq 2 \qquad (2)$$

Now, we have introduced the basic idea of calculating the number of configurations using a block with an α: α: α aspect ratio as a special case. Then we generalized the idea to estimate the possible number when the block is rectangular with the aspect ratio as α: β: β.



Appendix Fig 2. An illustration of the procedure to estimate the possible number of configurations when blocks have the aspect ratio of α: β: β. (a) Three types of rectangular blocks with an aspect ratio of α: β: β. (b) There are nine possible two-block configurations when combining blocks with an aspect ratio of α: β: β. (c) A three-block configuration could be viewed as stacking a cubic block on a two-block configuration.

The condition when the aspect ratio of blocks is α: β: β

A block with the aspect ratio of α: β: β has three types, corresponding to the sides of length, width and height are α and the rest sides are β (α: β: β, β: α: β, and β: β: α; see Appendix Fig 2a). For simplicity, we label the three basic blocks as A, B and C. The three types of blocks can generate 9 (i.e., 3^2) two-block configurations in total (Appendix Fig 2b). We calculate each of the possible numbers of two-block configurations below.

$$\boldsymbol{N_{R2}} = \begin{bmatrix} N_{AA} & N_{AB} & N_{AC} \\ N_{BA} & N_{BB} & N_{BC} \\ N_{CA} & N_{CB} & N_{CC} \end{bmatrix} = \frac{1}{\upsilon^2}\begin{bmatrix} 4\alpha\beta & (\alpha+\beta)^2 & 2\beta(\alpha+\beta) \\ (\alpha+\beta)^2 & 4\alpha\beta & 2\beta(\alpha+\beta) \\ 2\beta(\alpha+\beta) & 2\beta(\alpha+\beta) & 4\beta^2 \end{bmatrix} \quad (3)$$

The possible number of configurations for stacks containing two rectangular blocks with the aspect ratio of $\alpha{:}\,\beta{:}\,\beta$ is

$$N_{R2} = \sum \boldsymbol{N_{R2}} \quad (4)$$

For a configuration containing three blocks, it can be viewed as a block stacked on a two-block stack (Appendix Fig 2c). Therefore,

$$N_{R3} = N_{\cdot\cdot A} + N_{\cdot\cdot B} + N_{\cdot\cdot C} \quad (5)$$

Where $N_{\cdot\cdot A}$ indicates the possible number when block A stacked at the upper layer, and each term can be expanded as below.

$$\begin{aligned} N_{\cdot\cdot A} &= N_{\cdot A} \times N_{AA} + N_{\cdot B} \times N_{BA} + N_{\cdot C} \times N_{CA} \\ N_{\cdot\cdot B} &= N_{\cdot A} \times N_{AB} + N_{\cdot B} \times N_{BB} + N_{\cdot C} \times N_{CB} \\ N_{\cdot\cdot C} &= N_{\cdot A} \times N_{AC} + N_{\cdot B} \times N_{BC} + N_{\cdot C} \times N_{CC} \end{aligned} \quad (6)$$

Combining equations (4), (5) and (6), we have

$$N_{R3} = \sum \left([N_{\cdot A} \quad N_{\cdot B} \quad N_{\cdot C}] \times \begin{bmatrix} N_{AA} & N_{AB} & N_{AC} \\ N_{BA} & N_{BB} & N_{BC} \\ N_{CA} & N_{CB} & N_{CC} \end{bmatrix} \right)$$

And

$$[N_{\cdot A} \quad N_{\cdot B} \quad N_{\cdot C}] = [1 \quad 1 \quad 1] \times \begin{bmatrix} N_{AA} & N_{AB} & N_{AC} \\ N_{BA} & N_{BB} & N_{BC} \\ N_{CA} & N_{CB} & N_{CC} \end{bmatrix}$$

Therefore,

$$N_{R3} = \sum \left(\boldsymbol{N_{R2}}^2\right) \quad (7)$$

Following a similar logic, the possible number of configurations containing M blocks with an aspect ratio of $\alpha{:}\,\beta{:}\,\beta$ is

$$N_{RM} = \sum \left(\boldsymbol{N_{R2}}^{M-1}\right), M \geq 2 \quad (8)$$

The aspect ratio of blocks is $\boldsymbol{\alpha{:}\,\beta{:}\,\gamma}$

We further generalize the problem by considering the aspect ratio of blocks as $\alpha{:}\,\beta{:}\,\gamma$. This forms six different types: $\alpha{:}\,\beta{:}\,\gamma$, $\alpha{:}\,\gamma{:}\,\beta$, $\beta{:}\,\alpha{:}\,\gamma$, $\beta{:}\,\gamma{:}\,\alpha$, $\gamma{:}\,\alpha{:}\,\beta$, $\gamma{:}\,\beta{:}\,\alpha$, for each type the three proportional values corresponding to length, width and height, respectively. We label the six types of blocks as A, B, C, D, E, F, and G for simplicity.

Following the similar logic as above, different types of blocks generated 36 (i.e., 6^2) two-block configurations in total, and the possible number of each two-block configuration is

$$\boldsymbol{N_{R2}} = \begin{bmatrix} N_{AA} & N_{AB} & N_{AC} & N_{AD} & N_{AE} & N_{AF} \\ N_{BA} & N_{BB} & N_{BC} & N_{BD} & N_{BE} & N_{BF} \\ N_{CA} & N_{CB} & N_{CC} & N_{CD} & N_{CE} & N_{CF} \\ N_{DA} & N_{DB} & N_{DC} & N_{DD} & N_{DE} & N_{DF} \\ N_{EA} & N_{EB} & N_{EC} & N_{ED} & N_{EE} & N_{EF} \\ N_{FA} & N_{FB} & N_{FC} & N_{FD} & N_{FE} & N_{FF} \end{bmatrix} \quad (9)$$

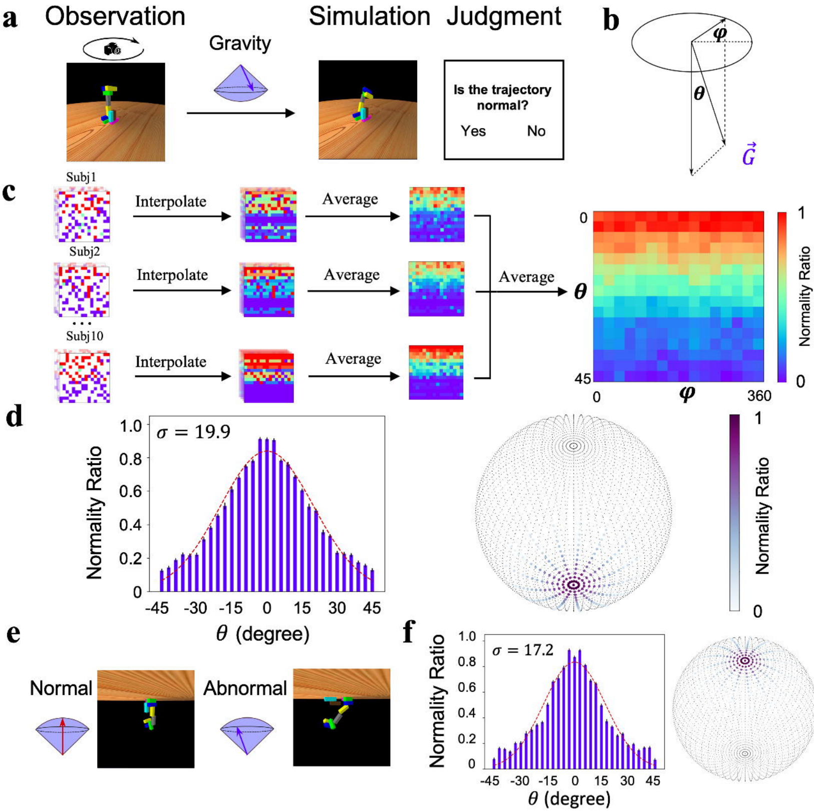
$$= \frac{1}{\upsilon^2}\begin{bmatrix} 4\alpha\beta & 2\alpha(\beta+\gamma) & (\alpha+\beta)^2 & (\alpha+\beta)(\beta+\gamma) & (\alpha+\gamma)(\alpha+\beta) & 2\beta(\alpha+\gamma) \\ 2\alpha(\beta+\gamma) & 4\alpha\gamma & (\alpha+\beta)(\alpha+\gamma) & 2\gamma(\alpha+\beta) & (\alpha+\gamma)^2 & (\alpha+\gamma)(\beta+\gamma) \\ (\alpha+\beta)^2 & (\alpha+\beta)(\alpha+\gamma) & 4\alpha\beta & 2\beta(\alpha+\gamma) & 2\alpha(\beta+\gamma) & (\alpha+\beta)(\beta+\gamma) \\ (\alpha+\beta)(\beta+\gamma) & 2\gamma(\alpha+\beta) & 2\beta(\alpha+\gamma) & 4\beta\gamma & (\beta+\gamma)(\alpha+\gamma) & (\beta+\gamma)^2 \\ (\alpha+\beta)(\alpha+\gamma) & (\alpha+\gamma)^2 & 2\alpha(\beta+\gamma) & (\alpha+\gamma)(\beta+\gamma) & 4\alpha\gamma & 2\gamma(\alpha+\beta) \\ 2\beta(\alpha+\gamma) & (\alpha+\gamma)(\beta+\gamma) & (\alpha+\beta)(\beta+\gamma) & (\beta+\gamma)^2 & 2\gamma(\alpha+\beta) & 4\beta\gamma \end{bmatrix}$$

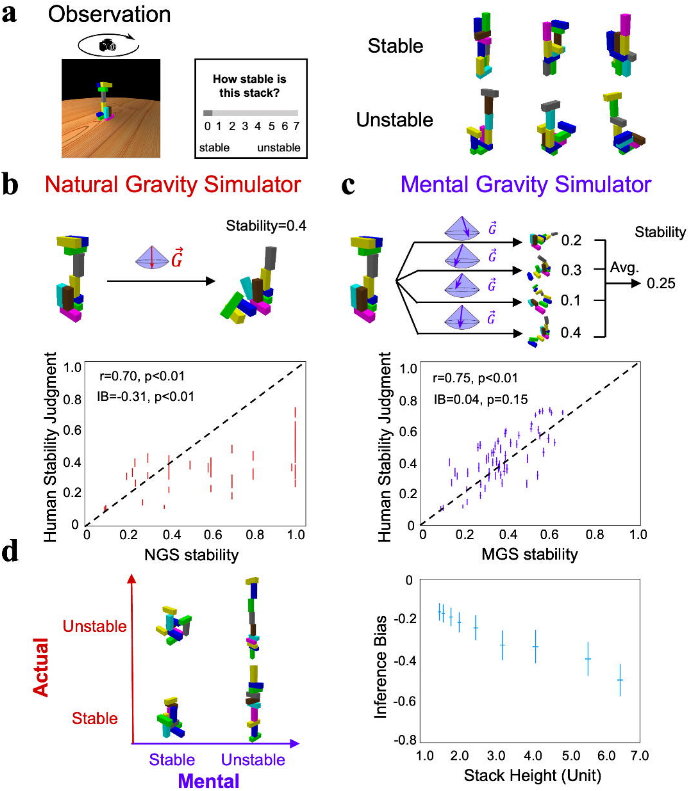
The possible number of configurations for stacks with M blocks with an aspect ratio $\alpha\!:\!\beta\!:\!\gamma$ is

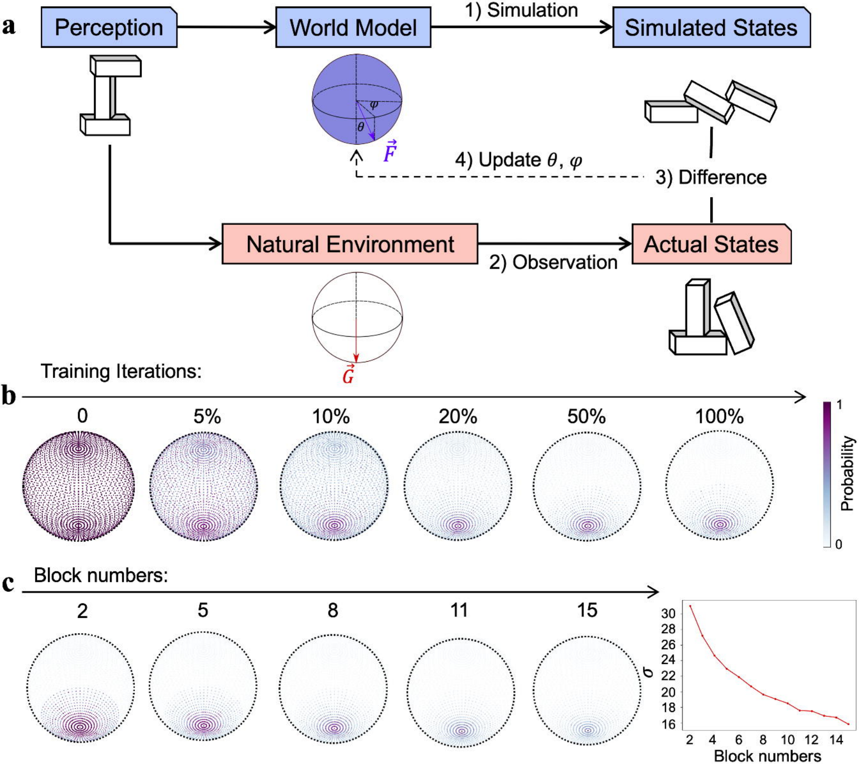
$$N_{RM} = \sum\left(\boldsymbol{N_{R2}}^{M-1}\right), M \geq 2 \tag{10}$$

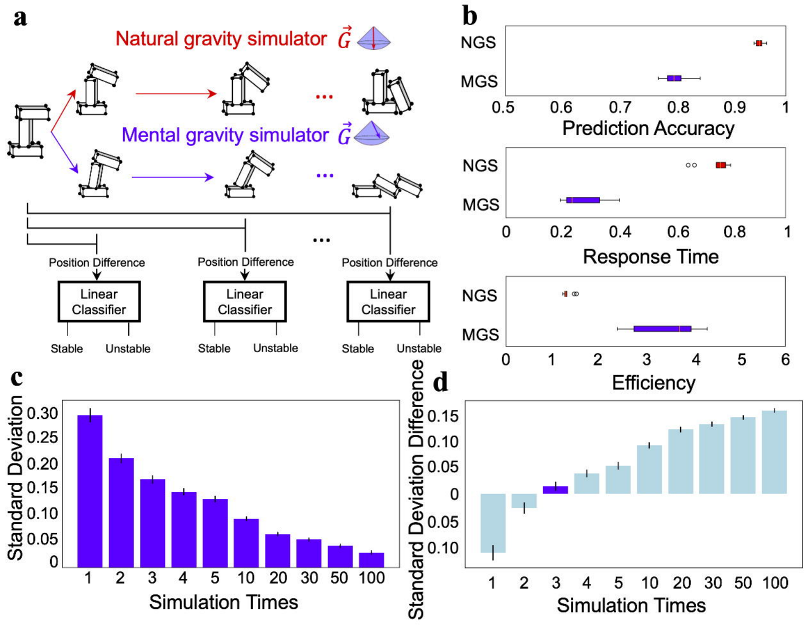
Therefore, we can estimate the possible number of configurations when only one rectangular block with the aspect ratio of $\alpha\!:\!\beta\!:\!\gamma$ is allowed to place in each layer using the formula (9) and (10).

Finally, in this study we chose blocks with an aspect ratio of 3:1:1 as building blocks for stacks whose stability was evaluated. Specifically, for stacks consisting of 10 blocks and j.n.d. of $\upsilon = 0.01$, the number of configurations can be estimated with formula (9), which is 3.72×10^{19}.


a
Observation
Gravity
Simulation
Judgment
Is the trajectory normal?
Yes
No
b
φ
θ
G
c
Subj1
Subj2
Subj10
Interpolate
Average
θ
φ
0
45
360
1
0
Normality Ratio
d
σ = 19.9
Normality Ratio
θ (degree)
e
Normal
Abnormal
f
σ = 17.2


a
Observation
How stable is this stack?
0 1 2 3 4 5 6 7
stable
unstable
Stable
Unstable
b
Natural Gravity Simulator
Stability=0.4
$\vec{G}$
c
Mental Gravity Simulator
Stability
0.2
0.3
0.1
0.4
Avg.
0.25
r=0.70, p<0.01
IB=-0.31, p<0.01
Human Stability Judgment
NGS stability
r=0.75, p<0.01
IB=0.04, p=0.15
MGS stability
d
Unstable
Stable
Actual
Stable
Unstable
Mental
Inference Bias
Stack Height (Unit)


a
Perception
World Model
1) Simulation
Simulated States
4) Update θ, φ
3) Difference
Natural Environment
2) Observation
Actual States
Training Iterations:
b
0
5%
10%
20%
50%
100%
Probability
1
0
Block numbers:
c
2
5
8
11
15
σ
Block numbers



a
Natural gravity simulator $\vec{G}$
Mental gravity simulator $\vec{G}$
Position Difference
Linear Classifier
Stable
Unstable
b
NGS
MGS
Prediction Accuracy
Response Time
Efficiency
c
Standard Deviation
Simulation Times
d
Standard Deviation Difference
Simulation Times