An Approximate Representation of Objects Underlies Physical Reasoning

Yichen Li[1,*] YingQiao Wang[1] Tal Boger[2] Kevin A. Smith[3]
Samuel J. Gershman[1] Tomer D. Ullman[1]

[1]Department of Psychology, Harvard University
[2]Department of Psychology, Yale University
[3]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology
[*]Corresponding author. E-mail: yichenli@fas.harvard.edu

Abstract

People make fast and reasonable predictions about the physical behavior of everyday objects. To do so, people may be using principled approximations, similar to models developed by engineers for the purposes of real-time physical simulations. We hypothesize that people use simplified object approximations for tracking and action (the *body* representation), as opposed to fine-grained forms for recognition (the *shape* representation). We used three classic psychophysical tasks (causality perception, collision detection, and change detection) in novel settings that dissociate body and shape. People's behavior across tasks indicates that they rely on approximate bodies for physical reasoning, and that this approximation lies between convex hulls and fine-grained shapes.

Keywords: intuitive physics | object representation | visual tracking | resource rationality | causality | collision detection | change detection

Introduction

Color, shape, and texture help us tell apples from oranges. But when trying to reason about an apple hurled towards your face, you may not care that it is green, or shiny, or even that it is an apple. All that matters is how fast, heavy, and elastic the apple is. For all reasonable purposes, it might as well be an orange.

We suggest that people use at least two representations of objects: *shape*, and *body*. The shape encodes features relevant for recognition, including fine-grain form. The body encodes properties relevant for tracking, collisions, and physical prediction. These properties include weight, position, and coarse form. The existence of something like a shape representation is not under dispute, though its exact nature has been greatly debated [1–3]. The existence of a body representation is a much less explored hypothesis, though across fields there are theories that people represent objects with limited fidelity. Here we propose a framework for why and how form representations should be limited when considering physical events.

The distinction between body and shape is motivated by engineering principles, and by converging evidence from cognition, developmental studies, and neuroscience. We next detail the relevance and convergence of these lines of research.

Engineers that design real-time physical simulators and game engines [4] often use principled approximations for greater speed and efficiency. Pressures of speed and efficiency may have led cognitive architectures to develop and adopt approximations similar to those used in such real-time simulators [5]. A central approximation used by real-time simulators is to approximate bodies for physical interactions, such as collision detection, separate from the fine-grain forms used for rendering objects. Body approximations can be refined meshes, but those are more computationally expensive, and approximations such as bounding boxes or convex hulls often produce reasonable results while reducing computational costs (see Figure 1).

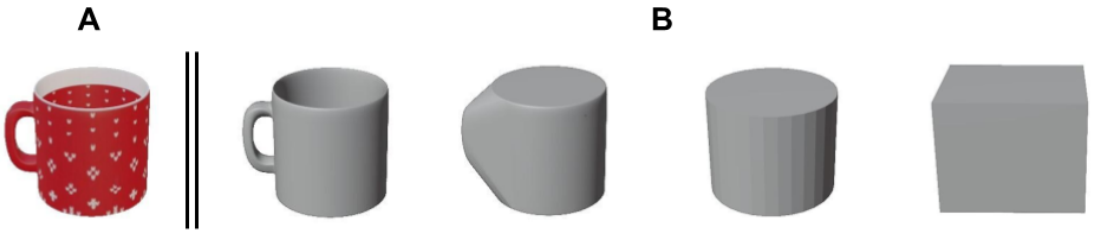

Figure 1: **Different representations for rendering and physical interactions**. **(A)** "Shape" is used for rendering an object onto the screen. **(B)** "Body" is an approximation used to determine collisions, apply forces, and track objects. Example approximations are shown in increasing coarseness from left to right: mesh collider, convex hull, cylinder collider, bounding box.

Previous work has proposed that noisy mental game engines underlie much of human intuitive physical reasoning [6–10]. This proposal has been challenged, with some researchers taking the mental game engine proposal to mean that intuitive physical reasoning should be a veridical simulation of reality. And, since physical reasoning deviates from reality, mental game engines are unsupported [11, 12]. However, it is likely that mental physical simulations (if they exist) use approximations in a resource-rational way, in line with resource-rational cognition [13, 14].

Studies in cognitive development show that infants below 12 months do not use fine-grained form information to track objects [15, 16], with follow-up work showing that such effects also exist in 18 months old under memory load [17]. These findings are often taken to suggest that young infants do not use "kind" information to track objects. We interpret them as showing that infants are relying on rough approximations for tracking. Such rough approximations are also central to recent artificial intelligence models that pass benchmarks designed to test models of core infant physics [18]. Other developmental work on change detection and occlusion has also led to a proposed distinction between features and objects in infant visual memory [19], which may map onto our body-shape distinction. If such a distinction exists early in

development, it likely persists into adulthood.

In neuroscience, a traditional split divides cortical visual processing in primates into ventral ("what") and dorsal ("where" or "how") streams [20, 21]. While the dorsal stream is often taken to encode spatial information about objects, more recent studies have refined this account [22], suggesting that the dorsal stream also encodes information that guides action. Research with non-human primates further suggests that the dorsal stream encodes action-relevant details of the form, orientation, and size of objects [23, 24]. Such an action-relevant form may in particular map on to a body approximation.

Taken together, findings from cognitive science, cognitive development, and neuroscience align with engineering principles to suggest that body approximations may be cognitively useful in physical reasoning, and separate from fine-grain forms for visual recognition. In order to examine the existence of this body-shape distinction in people, we created three distinct psychophysical tasks based on classic experiments: perception of causality in launching (Experiment 1), time-to-collision prediction (Experiment 2), and change detection (Experiments 3). While different in their design, the experiments shared an underlying logic, and used similar stimuli (see Figure 2A, and also Supplementary Information for details). We predicted that if people use body approximations for physical tasks, and given that these approximations partially fill a shape's concavity, then: (1) People would perceive collisions with concave objects as more causal than convex objects (Experiment 1), (2) People would predict concave collisions happen earlier than convex collisions (Experiment 2), and (3) People would be less likely to detect a change within the body approximation than an equally-sized change outside of it (Experiment 3). If people use the fine-grain shape for tracking, there should be no observable difference between concave and convex conditions across the experiments.

In addition to our general predictions, we considered a specific approximation model – α-shape – that allows us to quantitatively examine a space of possible body approximations [see also 25, and the Supplementary Information]. By changing one parameter (α), we examined body approximations from rough convex hulls, to fine forms (Figure 3 A). We do not take this specific model to be a process-level account of the approximation people use, as mathematicians and engineers have come up with many ways of simplifying and compressing shape information [4, 26]. Rather, the α-shape model allowed us to broadly differentiate between approximations closer to fine-grain forms, convex hulls, and intermediate representations.

Results

Experiment 1: Causality

Our first test of body approximations used causality judgments, based on the classic Michottean launching task [27]. Participants observed one shape (the Agent) moving towards another (the Patient). When the Agent stopped, the Patient started moving away from it (see Figure 2A, left). The horizontal spatial distance between the Agent and the Patient at the time when the Agent stopped moving and the Patient started moving varied from 0 to 64 pixels, where 64 pixels was about half the length of the Agent. Participants were asked to rate their agreement with the statement "The Agent caused the Patient to move". The Patient always had both a concave and convex side. On the concave side was a divot that would contain the point of contact should the two shapes collide. On the convex side, the collision point would be on the convex hull of the shape. The shapes were flipped horizontally across trials, allowing us to compare causality judgments for the same spatial gap and overall visual information, but varying whether the convex or concave side of the Patient was involved in the collision. The design, hypotheses, analyses, and exclusion criteria for this and all other experiments were preregistered (see Methods).

Michotte's original studies found that causality judgments decreased as the spatial gap at collision time increased [28]. We predicted that given the same spatial gap (i.e., the horizontal distance between the two objects at collision time), causality judgments in the concave condition would be higher than in the convex condition. The reasoning is as follows: Given that the body is coarse and fills in parts or all of a concavity, then if people used an approximate body to track the Agent and the Patient, the perceived spatial gap in

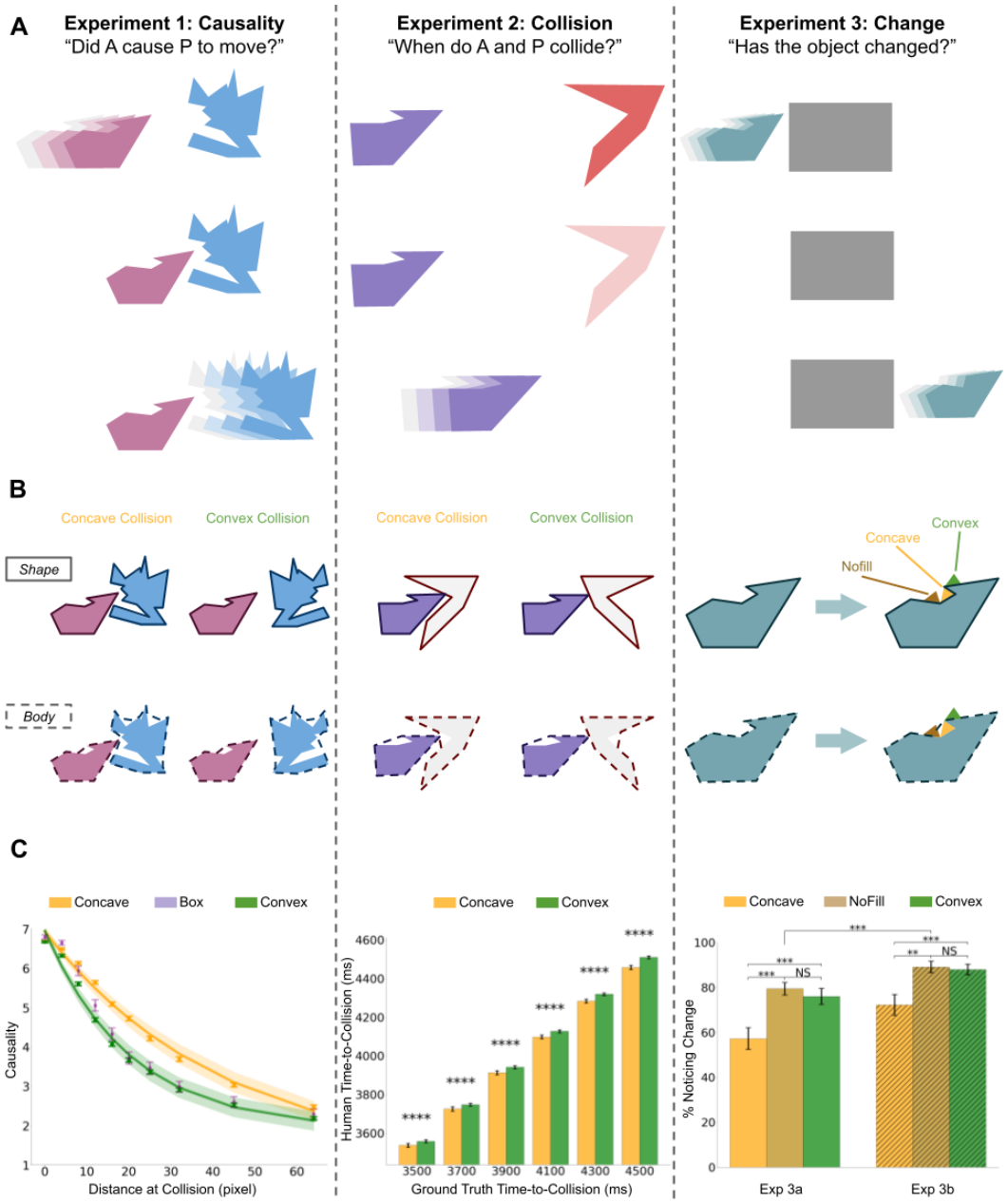

Figure 2: **Experimental design and results.** **(A)** Diagrams of stimuli **(B)** Differences in interactions between shapes and bodies. **(C)** Average participant responses (with SEMs and confidence intervals of curve fit). In **Experiment 1 (Causality, left)**, participants rated perceptions of causality when seeing an Agent (the first-moving object) colliding with a Patient (the second object to move). Coarse bodies result in a smaller perceived collision distance than concave collisions. We predicted and found that participants rated concave collisions as more causal than convex collisions, for the same spatial gap. In **Experiment 2 (Collision, middle)**, participants pressed a spacebar to indicate when an Agent and a (transparent) Patient collided. Coarse bodies result in smaller time-to-collision (TTC). We predicted and found that participants' TTC was smaller in concave vs. convex collisions, for the same ground-truth TTC. In **Experiment 3 (Change, right)**, an object changed or remained the same when passing behind an occluder, with the changes happening within or outside a coarse body approximation. We predicted and found that concave changes were more difficult to detect than changes outside the filled concavity, and changes outside the convex hull.

the concave condition should be smaller than the actual spatial gap. However, in the convex condition the perceived subjective spatial gap should be close to the actual spatial gap, as the body approximation does not differ much on the convex side of objects compared to the original shapes (imagine a convex hull being used as the approximate body).

We fit two exponential decay curves to people's causality ratings as a function of the horizontal distance during the moment of collision, one curve for convex collisions and one for concave collisions (see 2C, left). Specifically, we used $C = a * e^{-D/b} + c$, where C was the rating of causal perception, a, b, and c were free parameters, and D was the horizontal collision distance (the spatial 'gap'). We predicted that under the same collision distance, concave trials would be perceived as more causal than convex trials, meaning the concave curve would be mostly above the convex curve, but may converge at the edges. Specifically, we predicted $b_{concave} > b_{convex}$, showing a significant difference of the curvatures between the two curves while parameters a and c might constrain the two curves' ending points to overlap. Using least squares fitting, we found that the best parameters were (with 95% CI in parentheses): $a_{concave}$ = 5.8 [5.3, 6.0], a_{convex} = 5.04 [4.7, 5.3]; $b_{concave}$ = 41.6 [35.6, 47.1], b_{convex} = 20.6 [18.3, 22.9]; $c_{concave}$ = 1.1 [1.0, 1.5], c_{convex} = 1.9 [1.7, 2.1]. A paired t-test on the bootstrapped $b_{concave}$ and b_{convex} showed that indeed the two are significantly different ($T(999) = 3.1 \times 10^2$, $p < 0.001$); 100% of 10000 bootstrapped comparisons showed that bootstrapped $b_{concave} > b_{convex}$.

As predicted by a body-shape distinction, concave collisions were perceived as more causal than convex collisions given the same ground-truth spatial distance at collision time. A control experiment further examined the possibility that participants were using the Euclidean distance between objects rather than the horizontal collision distance between them. The control experiment used shapes where the Euclidean distance and the horizontal distance were disentangled. The results replicated the main effects of Experiment 1, and further demonstrated that Euclidean distance did not account for people's ratings (see Supplementary Information).

We used an α-shape approximation algorithm [25] to examine in more detail the approximation people may be using. The α-shape algorithm produces an approximate polygon of a given shape, with one parameter α controlling the coarseness of the resulting approximation, ranging from a convex hull to a fine-grain form (Figure 3A). Each setting of α produced different predictions of the perceived collision distance between the Agent and the Patient. The best-performing α value among the ones tested (ranging from convex hull to shape) was 0.051 (Figure 3B, left). The corresponding average area-difference percentage between the approximate body and the original shape is 21.9%. This model accurately explained participant causality ratings for both concave and convex conditions (Figure 3C, left). This parameter setting aligns with a body approximation that is between a convex hull and a fine-grained shape, further supporting the hypothesis that body is not equal to shape.

Experiment 2: Collision

Experiment 2 was based on classic time-to-collision tasks [29–31], and tested people's predicted collision time between two objects. Participants saw 4-5 second videos of two objects, an Agent and a Patient (Figure 2, middle). At first, Agent and Patient were stationary. The Patient then faded away, and the Agent began moving at constant speed towards the now-invisible Patient. Participants were asked to press the spacebar at the moment of Agent and Patient collision. The time between the Agent launch and the participant's press of the spacebar was coded as time-to-collision (TTC). We manipulated the initial horizontal distance between the Agent and the Patient, creating 6 ground-truth TTCs varying from 3500ms to 4500ms, which corresponds to a different of between 1.6 to 2.5 the length of the Agent. Similar to Experiment 1, we used mirror images of each Patient shape to create concave and convex contrasts. See the Supplementary Information for more details.

We predicted that people's TTC would be based on the approximate bodies of the Agent and the Patient, in which case people's TTC for concave collisions should be smaller than convex collisions with the same ground-truth TTC. This is because a coarse body representation partially fills in shape concavities, making

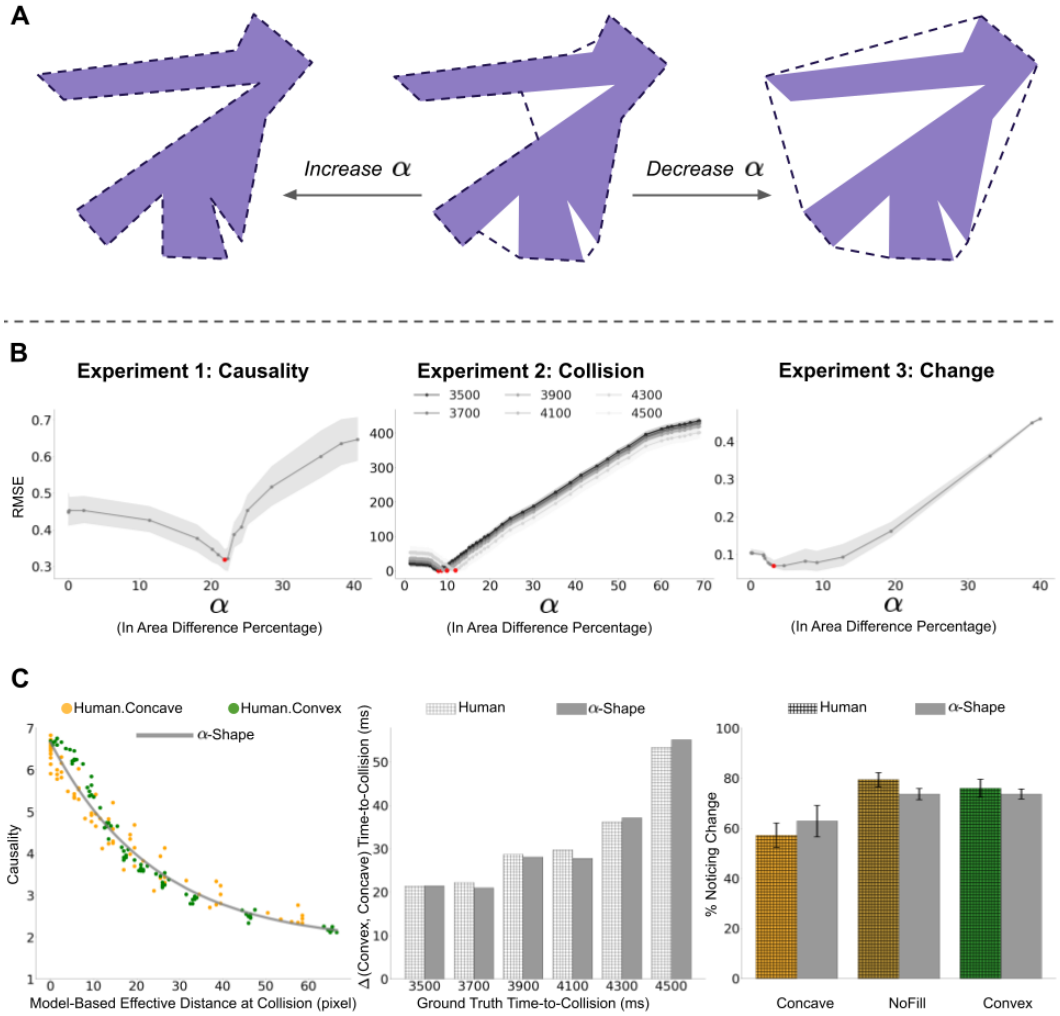

Figure 3: **α-shape model overview, and modeling results**. **(A)** The α-shape model (dotted lines) fits different approximations to a given shape (solid purple). As α increases, the approximation overlaps more with the original shape, and as α decreases, the approximation is closer to a convex hull. We chose the range of α values to cover approximations from convex hulls to fine forms. **(B)** Best-fit α-shape models in the different experiments. The x-axis shows the α parameter using average area-difference percentage between the original shape and the approximated body (raw α values are non-linear and less informative), with the left-most point (0) corresponding to body = shape, and the right-most point corresponding to body = convex hull. The y-axis shows root-mean-square-error (RMSE) and 95% confidence intervals when using different α values to predict participant responses (lower values suggest a closer fit). The best α-shape models (i.e., with the lowest RMSE) are indicated by a larger red dot. Experiment 1 used single-value α-shape models, Experiment 2 used both time-varying α-shape models (shown here) and single-value models, and Experiment 3 used models taking into account shape complexity. **(C)** Prediction from the best-performing α-shape model in each experiment compared with participant data.

the perceived distance that the Agent must travel to contact the Patient shorter in concave collisions.

We compared the distributions of participant TTCs in concave and convex collisions across all ground-truth TTC conditions using Kernel Density Estimation. The average difference (and the 95% confidence interval) between participant concave TTC and ground-truth TTC (i.e., $\Delta TTC_{concave}$) was 2.4ms [-1.4, 6.2]; the average difference between participant convex TTC and ground-truth TTC (i.e., ΔTTC_{convex}) was 34.3ms [31.8, 36.7]. As expected then, people overall performed as if concave collisions happened earlier than convex collisions, given the same ground-truth TTC.

We next considered each ground-truth TTC seperately, and tested if the difference between concave and convex TTC still held. We found that participant concave TTC and convex TTC were significantly different in every ground-truth TTC condition (Figure 2C, middle): $T(1161) = -4.4, p = 1.3 \times 10^{-5}$ in 3500ms ground-truth TTC condition; $T(1181) = -4.4, p = 1.2 \times 10^{-5}$ in 3700ms ground-truth TTC condition; $T(1182) = -5.7, p = 1.2 \times 10^{-8}$ in 3900ms ground-truth TTC condition; $T(1179) = -5.9, p = 3.9 \times 10^{-9}$ in 4100ms ground-truth TTC condition; $T(1133) = -7.2, p = 8.6 \times 10^{-13}$ in 4300ms ground-truth TTC condition; and $T(1145) = -10.3, p = 2.2 \times 10^{-16}$ in 4500ms ground-truth TTC condition. In short, participants predicted concave collisions happened earlier than concave collisions regardless of the ground-truth TTC. The results again align with our hypothesis of a body-shape distinction.

An exploratory analysis further found that as the ground-truth TTC increased, the difference between participants' convex TTC and concave TTC also increased. This is in line with a memory effect on the coarseness of the approximation, such that the body approximation grows coarser in working memory over time.

As in Experiment 1, we also used an α-shape model with different values of α to produce different approximations of the shapes used. The approximations produced different corresponding TTCs. In this experiment we considered both a static α (see Supplementary Information) and a time-varying α, which corresponds to the memory-effect on approximation coarseness. If the body became coarser over time, presumably more area in the object's concave divots was filled in, resulting in a larger error in participant concave TTC compared to the ground-truth TTC. Thus, we allowed the α value to vary across ground-truth TTC conditions. We found suggestive evidence that a time-varying α-shape model better fit participant data than a single-value (static) α-shape model. The best α value in each ground-truth TTC condition (3500ms to 4500ms) was 0.044, 0.044, 0.042, 0.042, 0.038, and 0.032, suggesting the approximation is growing coarser over time. These best α values filled in the concave divot partially for about 7.9% to 11.9% in size with respect to the original shape (Figure 3B, middle). The best time-varying α-shape parameters reproduced the memory effect in participant data (Figure 3C, middle), such that the difference between convex and concave TTC increased over time. This again supports a body approximation that is in between a convex hull and the shape.

Experiment 3: Change

Experiment 3 was based on classic change-detection findings [32, 33]. In Experiment 3a, we adopted the infant change detection paradigm from [15], in which an object (e.g., a duck) moves behind an occluder, and another object (e.g., a truck) emerges. Infants at around 10 months of age are at chance at predicting the toy duck will be behind the occluder after the truck has exited. This type of paradigm is often used to explore object individuation and identity in infants [34, 35], and has been taken to suggest young infants do not use 'kind information' to track objects. In previous work, we suggested that this finding may be due to a body-shape distinction, with the two objects having roughly similar bodies [5].

In Experiment 3a, participants watched short videos of an object moving behind an occluder (Figure 2A, right). After the object was fully occluded, another object emerged, either identical to the first, or differing by some added area. Participants had to decide whether the object that exited the occluder was the same as the object that entered it. We predicted that if people use approximate bodies for physical tracking, then participants would notice changes at a higher rate when the added area caused larger changes to the body representation of the object.

We tested our hypothesis by adding area to a given shape either in an inner concavity (*concave* condition), a non-inner concavity (*nofill* condition), or a convexity (*convex* condition), and see Figure 2B, right. We also created two sizes for the added area in each condition. If the body representation is the same as the shape, then changes in the concave, nofill, and convex conditions should be noticed at the same rate. If the body approximation is a convex hull, nofill and concave changes should be equally harder to detect than convex changes, because concave and nofill changes happened within a convex hull. If the body approximation is somewhere between a convex hull and a fine-grained shape (as suggested by Experiments 1 and 2), then the concave changes should be harder to detect than the nofill and convex conditions, because only concave changes would fall within the body.

As shown in Figure 2C (right), we found that indeed the odds that participants noticed a change were lowest for the concave trials, significantly lower than either nofill trials (sample mean difference was 22.2%, $t(55) = 9.49, p < 0.001$; $d = 1.27$; 95% CIs = [17.5%, 26.9%])) or than convex trials (sample mean difference = 18.7%, $t(55) = 7.29, p < 0.001$; $d = 0.97$; 95% CIs = [13.5%, 23.9%]. The difference between the nofill and convex trials was not significant (sample mean difference = 3.4%, $t(55) = 1.76, p = 0.084$; $d = 0.24$; 95% CIs = [-0.4%, 7.3%]). This suggests that the body approximation is not equal to shape, and further that the boundaries of body does not contain the changes in the nofill condition, in line with Experiments 1 and 2.

We followed Experiment 3a with a control, in which no direct physical-tracking of motion was involved. In this Experiment 3b, we used static images as stimuli. Participants watched a stationary object at the center of the screen for 1 second, after which the object disappeared for 2 seconds (this period matched the approximated time that the object was hidden behind the occluder in 3a). The same object or a modified object then appeared for another 1 second. As in Experiment 3a, we created three change conditions (concave, nofill, and convex), and two sizes of the change. We hypothesized that in Experiment 3b, concave trials would be easier to detect, because the absence of physical movement would result in less dependence on a body representation. The findings replicated the overall pattern of Experiment 3a, but with change-detection being easier across the board (Figure 2C, right). A two-way logistic ANOVA showed that the interaction between change type (concave, convex, or nofill) and experiment version (3a or 3b) was not significant, and both main effects of change type and experiment version were significant (interaction: $\chi^2(2)$=1.08, p=.58; change type main effect: $\chi^2(2)$=211.60, p<.001; experiment version main effect: $\chi^2(1)$=112.45, p<.001). This suggests that the visual task in 3b was easier than the physics-tracking task in 3a, but without a differential effect on detecting concave changes. It is possible that having the before- and after-image presented sequentially but with a temporal gap, caused people to maintain a coarse body-like representation in memory in order to perform visual comparison.

Finally, and as in Experiments 1 and 2, we examined different values of α in our α-shape model for Experiment 3a, ranging from a convex hull to a fine form similar to the original shape. For each α, we used the effective area change between the body approximations of the shapes before and after the change to predict the percentage of of noticing a change. We also took into account the visual complexity of the original shape, such that changes to more complex shapes were predicted to be more difficult to detect regardless of body approximations (see the Supplementary Information). The best-performing overall average α fills in on average of 4.6% of the concavities of the original shape. We stress that this should not be taken to suggest that the true underlying body approximation fills in concavities to this specific amount, but simply that the approximation fills in the concavities to some degree, and further work should elucidate the specific approximation used.

Discussion

Our results suggest that people use a coarse approximation for reasoning about the behavior of objects, and that this body representation is in between a convex hull and a fine-grained form. Such a body approximation is in line with the general proposal that people's intuitive physics is not a perfect simulation, but rather relies on principled short-cuts and workarounds [5, 6, 8, 36]. It also supports the proposal that

cognitive scientists can use the principled approximations of real-time simulations as working hypotheses for cognitive models of intuitive physics. Other approximations to explore include wake-sleep, and static-dynamic distinctions [5]. While this set of experiments provides a first step in showing that people use body approximations for reasoning about physical events, further work is required to determine when people use approximate body representations, how they are formed, and how they might change across time and tasks.

It is likely that people's approximations are task- and context-specific in a dynamic way. The simple α-shape model we considered treated all parts of the shape as equally important, whereas it is quite possible that people use less resources for areas of the shape that are less relevant. For example, if an object is about to experience a collision on its left-hand side, it may be unimportant to spend resources on approximating its right-hand side. The importance and difficulty of the task may also affect the approximation used [37]. For example, if it is vitally important to precisely assess the trajectory of a object, more cognitive resources may be spent on finer-grained approximations to increase accuracy. The results of Experiment 2 also suggest the approximation model is time-variant, with people's approximation growing rougher with time up to a point (the body approximation may grow closer and closer to a convex hull the longer it spends behind an occluder or in memory). All of these suggestions are not alternatives to the current proposal, but rather suggestions for refinement. These suggestions also easily lend themselves to further experiments, and additions to the model.

The α-shape model we considered is useful in teasing apart several possibilities for whether and which approximation people use, but it is only one suggestion for the approximations people might use when simplifying two-dimensional shapes, based on [25]. It's quite likely that people do not use exactly this model. Various shape-simplification models have been put forward by mathematicians, and possibly different algorithms are used for 2D vs 3D approximations [38]. Follow-up work can further constraint the different approximation model(s) used by people.

Body approximations may also be influenced by kind-information. For example, a cylinder may be used to approximate a mug, but it is important for a prototypical mug that it has a handle. Such information is useful for recognition, but also for making physical predictions. A useful body-approximation algorithm may include a library of standard shapes [cf. 18] that is expanded over time, with language helping to scaffold the importance of different shapes. The failure of infants to detect a change in shape when objects move behind an occluder [15] may then reflect either a very rough body approximation, or the lack of relevant bodies in a standard body library.

While kind-information may help constrain body approximations, this can only happens up to a point, and some insensitivity to kind-information may carry through from infancy to adulthood. For example, it was recently shown [39] that people "fill in" the perceived trajectory of objects, even when those objects change identity (from a basketball to a soccer ball). But, this effect did not exist when objects changed spatio-temporal continuity (a basketball is seen coming in from above, then from below). Our proposal predicts such behavior, and further predicts that changing the object outside of a rough body approximation will disrupt filling-in effects (for example, changing a basketball to a much larger basketball, or a basketball to a towel).

Returning to the dorsal-ventral distinction in visual processing in primates [20–22], a body-approximation would be in line with information-for-action, rather than recognition. Above and beyond "where" something is, acting on something requires knowing its rough physical form. A small doughnut centered in a particular position is not the same as a large box centered on the same location. In game engines, the body-representation is a carrier not just of rough form, but also of orientation, location, and physical properties like elasticity and weight. It is an interesting avenue for future research, to examine to what degree this analogy carries into primate visual processing, although it is unlikely to be a neat split [40].

Moving from perception to imagery, the split between "visual" and "spatial" imagery has been noted previously [e.g. 41]. It has also been considered in more detail recently in research on aphantasia [42, 42–45], which refers to some people's inability to form voluntary visual images. Many of these individuals are able to pass tasks considered the domain of visual imagery, such as mental rotations. Our results suggest that body-representations may be the relevant forms preserved in spatial imagery. To use a crude analogy

with physics engines, aphantasia may be an issue with the "rendering" operation, while other computations are intact. This is similar to how one can run a physical simulation, without rendering a scene on a screen.

In sum, our findings suggest that human perception and reasoning respects the body-shape distinction. We used a contrast between concave and convex trials in three psychophysical tasks to create a dissociation between body and shape. We observed in all three experiments that human behavior in concave trials was significantly different from convex trials, as predicted by a distinction between body and shape representations. We used the α-shape algorithm to produce a specific realization of body representations, and found that reasonable α values predict human behavior. These specific α values all distinguish body from shape. While our models are unlikely to be a perfect match for people's representations, our finding suggest they are a decent approximation.

Methods

In-depth details describing the experiments and the α-shape model can be found in the Supplementary Information. Experiments – including design, hypotheses, analyses, exclusion criteria – were preregistered at `https://osf.io/f3kwd` (Experiment 1), `https://osf.io/unfzd` (Experiment 2), `https://osf.io/krzq2` (Experiment 3a), and `https://osf.io/nre7s` (Experiment 3b). Data, stimuli, and analysis code for all experiments are openly available at `https://osf.io/z9dpu/`. Experiment 1 tested people's perception of physical causality using Michottean launching with concave and convex shapes, varying the collision distance. Experiment 2 tested predicted collision times between two objects, varying concave and convex collision types and the ground-truth collision times. Experiment 3a tested change detection for a shape moving behind an occluder, where a change happened within or outside a potential body approximation; Experiment 3b controlled for the physical motion in Experiment 3a.

Participants

Sample sizes were determined by power analysis (99% power, significance level = 0.05) based on pilot data for each experiment, with the exception that Experiment 3b used the same number of participants to match 3a. A total of 670 participants were recruited online, through Amazon Mechanical Turk [Experiment 1; 46] or through Prolific [Experiments 2 & 3; 47].

In Experiment 1, 330 participants (female = 84; median age = 37; median completion time = 26.2 minutes) were recruited online through Amazon's Mechanical Turk service, with a link directing to a survey page on Qualtrics. Participants were compensated 4.5 USD for their time, at a rate of about 10 USD per hour. After applying the exclusion criteria, 147 participants were left for analysis. Experiment 2 recruited 226 participants through Prolific. After applying the exclusion criteria, 178 participants were left for analysis (women = 88, men = 83, non-binary = 2, declined to answer = 5; median age = 32; mean completion time = 20 minutes). We recruited 60 participants each for Experiments 3a and 3b through Prolific (average completion time = 20 minutes). In Experiment 3a, after excluding invalid data, we were left with 56 participants. In Experiment 3b, 50 participants remained after applying the same exclusion criteria. For both Experiments 3a and 3b, demographic data was optional and the majority of participants chose not to provide it. See exclusion criteria for each experiment in Supplementary Information.

All participants were US-based, and all experiments were approved by the Harvard University Area Institutional Review Board (protocol no. 19-1861)

Procedure and Stimuli

Figure 2A shows example sketches of stimuli used in the experiments. Screenshots from the experiments and links to stimuli can be found in Supplementary Information. Across all experiments, the stimuli was based on the same basic set of 8 irregular shapes, taken from a classic study on mental rotation [48] for having low verbal association.

Experiment 1 was based on the classic Michottean launching task [27, 49], which tests the perception of physical causality. Participants saw 5-second videos of an object (the Agent) moving towards a stationary object (the Patient). At a pre-determined point, the Agent stopped moving, and the Patient began moving away from the Agent, in the same speed and direction as the Agent did (Figure 2A, left). At the end of each video, participants used a 7-point Likert scale to report their agreement with the statement "The Agent caused the Patient to move" [cf. 49]. This level of agreement was the dependent variable. The horizontal distance between the Agent and the Patient at the time of collision was one of the following values: 0, 4, 8, 12, 16, 20, 25, 32, 45, and 64 pixels. The longest distance, 64 pixels, corresponded to about half the length of the Agent. The Patient always had both a concave side with a divot that would contain the point of collision if the two objects were to contact, and a convex side. We created mirror images of the 8 irregular shapes, flipped and applied a slight rotation to align with the Agent's point of contact. Concave trials had the Agent moving towards the concave side of the Patient (i.e., the point of contact was within the concave divot of the Patient), and convex trials had the Agent moving towards the convex side of the Patient (i.e., the point of contact was on the convexity of the Patient). This allowed us to compare participants' causality judgments between concave and convex side-of-hit, fixing a spatial distance and the overall visual complexity. We randomized the direction of motion (left to right, or right to left) and the Agent/Patient colors in every video. We used either irregular shapes or box-shapes as the Agent and Patient. For trials with irregular shapes, the Agent was always the same shape, while the Patient shape varied across trials. Trials with box-shapes served as warm-ups and controls. In total, there were 180 videos (10 distances x 8 shapes x 2 side-of-hit conditions + 20 regular box trials). Participants first saw 10 warm-up collisions with regular boxes to establish baselines and exclusion criteria, followed by a randomized presentation of the other 170 videos.

Experiment 2 was based on time-to-collision tasks [29–31], and tested people's prediction of the collision time between two objects. Participants saw 4-5 second videos of two objects starting stationary. After 1.6 seconds, the Patient faded, and the Agent began to move horizontally towards the Patient's last location. Participants were instructed to press the spacebar at the time when they predicted the Agent and the Patient collided. The time difference between the Agent initiating motion and the spacebar-press was the dependent variable, time-to-collision (TTC). We used different horizontal distances between the Agent and the Patient, corresponding to 6 ground-truth TTCs: 3500ms, 3700ms, 3900ms, 4100ms, 4300ms, and 4500ms. As in in Experiment 1, we created concave and convex conditions for every irregular shape by varying their side-of-collision. We lightly simplified the shapes from Experiment 1 to help with the brief-timing of motor responses. Specifically, we made the concavities on objects larger so that the difference (if there was any) in response times would be comparable between concave and convex conditions. We used either irregular shapes or box-shapes as the Agent and Patient. 24 control videos were inserted in the experiment, in which the Patient did not vanish and remained visible throughout the video. Box-shapes and non-fading Patients were used to establish baselines, ceiling performance, and exclusion criteria. We randomized the horizontal trajectory of the Agent (either left-to-right or right-to-left) and colors of objects in each video. In total, we had 120 videos (96 test videos: 6 ground-truth TTC conditions x 8 irregular shapes x 2 side-of-collision conditions; 24 control videos: 20 with irregular shapes but no vanishing + 2 with boxes and vanishing + 2 with boxes and no vanishing).

Experiment 3 was based on classic change-detection tasks [32, 33], and on studies with young children in which shape-changes behind occluders are not noticed in early ages [15–17]. This experiment tested people's ability to detect whether a change occurred in an object's shape. In Experiment 3a, Participants saw 4-second videos in which an object moved horizontally behind a centrally placed occluder. The object was briefly out of sight when it moved behind the occluder, and then either the same object or a modified object emerged. Modified objects had areas added to them, in three locations (concave, nofill, and convex) and two sizes (small and large). The concave condition filled an inner-most concavity of an object, the nofill condition had the added area still within a big concavity, but not necessarily filling in the inner-most position, and the convex condition had the added area at a convex edge of the object. The pixel-area change and form of the added area was the same across change types, within a size and shape. Participants reported

whether or not they detected a change to the object after viewing each video, and this binary measure was our dependent variable. The base objects were the same 8 irregular shapes used in Experiment 1. We balanced the number of videos showing change and no change, and randomized the horizontal motion of the object (either from left to right, or from right to left) and well as its color. In total, there were 96 test trials (8 shapes x 3 change types x 2 change sizes x 2 change/no-change conditions). We also created catch trials as attention checks, in which a simple square stayed as a square, or changed into a triangle. Experiment 3b was a control for physical motion in Experiment 3a. Experiment 3b replicated the design and measures of Experiment 3a, except that we used static images instead of videos of moving object. The object was stationary at the center of the screen (1 second), then disappeared (2 seconds) and the same object or a modified object appeared (1 second). Timings were based on conservative estimates of how long each object was on the screen in Experiment 3a.

Data Analysis

In Experiment 1, we fit two separate exponential decay curves to participant causality judgment ratings in concave trials and in convex trials separately, with the formulation $C = a \cdot e^{-D/b} + c$. The curves predicted participant causality ratings (C) using the ground-truth horizontal collision distance (D) as input. Free parameters $a, b,$and c were estimated by least squares optimization. Parameters a and c controlled the displacement and intercept of the curve, and had an effect on converging the starting- and ending-point of the two curves. We were interested in the curvature of the two curves, namely the difference between $b_{concave}$ and b_{convex}. We used 1000 bootstraps of participant responses. In every bootstrap, we sampled the full sample size with replacement, averaged responses across sampled participants for every ground-truth collision distance, and fit curves over the averaged data. In total we obtained 1000 bootstrapped fitting results of every parameter in the curve. A paired t-test was used to compare the 1000 bootstrapped $b_{concave}$ fits and the 1000 bootstrapped b_{convex} fits. We repeated this comparison between $b_{concave}$'s and b_{convex}'s 10000 times. In every repeat, we sampled 1000 $b_{concave}$'s and b_{convex}'s with replacement, and tested their distribution by a paired t-test (α=0.05). We calculated the percentage of repeats with significant t-test results. Finally, we did an exploratory α-shape analysis. By varying the α value, we obtained a range of α-shape models to cover approximations from convex hulls to a fine-grained forms (although the absolute magnitude of the α value range may vary across tasks, depending on the scale of images). Every setting of α produced approximated representations for the Agent and the Patient. We used these approximations to estimate the effective collision distance, i.e., the horizontal distance between the two approximations at the collision moment. Then, using the effective collision distance as input, we fit a single exponential decay curve (same formulation as above, but replacing D with the effective collision distance) with least squares to predict participant causality ratings for both concave and convex trials at once. The expectation was that a reasonable α-shape model should produce body approximations that can account for causality ratings in both concave and convex trials, removing their non-overlap. The performance of the fit was measured by root mean squared error (RMSE) between each participant's response and the model's prediction. We reported the best α-shape model with the least RMSE.

In Experiment 2, we first compared concave and convex conditions by aggregating across all participants and all ground-truth TTCs. We used Kernel Density Estimation with a Gaussian kernel to estimate the ΔTTC distribution for concave and convex trials separately. The ΔTTC is defined as the difference between participant TTC and the ground-truth TTC. We then compared the mean and 95% confidence interval between the estimated $\Delta TTC_{concave}$ distribution and the ΔTTC_{convex} distribution. Next, we considered each ground-truth TTC condition. We compared the TTC in concave trials and in convex trials for every ground-truth TTC condition using paired t-tests (α=0.05). Finally, we used an exploratory analysis to test different settings of the approximation parameter in an α-shape model ranging from convex hulls to fine-forms. Every α produced approximated representations for the Agent and the Patient. We estimated the effective TTC between the two approximations given the ground-truth speed of the Agent. We then used a hierarchical linear model to predict participant TTC responses by using the effective TTC as input. The α-shape model

performance was again measured in RMSE. Here, we allowed a time-varying α-shape model to account for the memory effect we observed in participant responses (that the difference between convex TTC and concave TTC became larger as ground-truth TTC increased.) In other words, we allowed the best α value in each ground-truth TTC condition to vary. We report the best-fitting α-shape model and the best α value for each ground-truth TTC. See more details of the α-shape modeling and a single-parameter comparison in the Supplementary Information.

In Experiment 3a and 3b, we calculated the average percentage of detecting change, and the SEM in each change type condition (i.e., concave, nofill, and convex), using only the data from change trials and aggregating across change sizes. We compared the percentage of noticing change among change types by paired t-tests (α=0.05). Next, we performed a generalized linear regression with logistic link function (i.e., the binomial family) on participant data from both Experiment 3a and 3b. The parameters included a main effect of change type (concave, nofill, convex), a main effect of experiment version (3a, 3b), and their interaction. We also performed ANOVA χ^2 tests on the regression model, and report its deviance and significance level. Finally, we tested different α values for an approximation model ranging from convex hulls to the fine-form of the original shapes. Each α setting produced an approximation for the objects before and after a change. We aligned the two approximations, and calculated the proportion of area that was different in the two approximations with respect to the size of the before-change approximation (i.e., the relative and effective amount of body violation). Independently of the approximation, we took complexity of the original shape into account, as we found that the odds of noticing a change varied across shapes (which themselves varied in complexity). We parametrized visual complexity as the number of vertices a given shape has before entering the occluder. We used the effective area change ratio x and *complexity* to predict the percentage of change detection $P(change)$, with the logarithm functional form $P(change) = (P(falseAlarm)+a)*log(e+b*x)-a+k*complexity$. Free parameters a, b, and k were estimated using least squares. The constant $P(falseAlarm)$ was the false alarm rate of participants reporting change in the no-change trials containing irregular shapes. Performance was measured using the mean RMSE across averaged concave predictions, averaged nofill predictions, and averaged convex predictions. We report the best α-shape model with the least RMSE. See more details on α-shape modeling in Supplementary Information.

Acknowledgments

We thank Joshua Tenenbaum for helpful comments, and Eric Bigelow for discussions and analysis. T.D.U. and K.A.S. are supported by NSF Science Technology Center Award CCF-1231216, and the DARPA Machine Common Sense (MCS) program. T.D.U is supported by the Jacobs Foundation Fellowship. K.A.S. is supported by NSF grant: Adversarial Collaborative Research on Intuitive Physical Reasoning (#2121009).

References

[1] David Marr. *Vision: A computational investigation into the human representation and processing of visual information*. Henry Holt & Company, 1982.

[2] Shimon Ullman. Aligning pictorial descriptions: An approach to object recognition. *Cognition*, 32(3):193–254, 1989.

[3] Irving Biederman. Recognition-by-components: a theory of human image understanding. *Psychological review*, 94(2):115, 1987.

[4] Jason Gregory. *Game engine architecture*. crc Press, 2018.

[5] Tomer D. Ullman, Elizabeth Spelke, Peter Battaglia, and Joshua B. Tenenbaum. Mind games: Game engines as an architecture for intuitive physics. *Trends in Cognitive Sciences*, 21(9):649–665, sep 2017.

[6] Peter W Battaglia, Jessica B Hamrick, and Joshua B Tenenbaum. Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, 110(45):18327–18332, 2013.

[7] Adam N Sanborn, Vikash K Mansinghka, and Thomas L Griffiths. Reconciling intuitive physics and newtonian mechanics for colliding objects. *Psychological review*, 120(2):411, 2013.

[8] Kevin A Smith and Edward Vul. Sources of uncertainty in intuitive physics. *Topics in cognitive science*, 5(1):185–199, 2013.

[9] Jessica B Hamrick, Peter W Battaglia, Thomas L Griffiths, and Joshua B Tenenbaum. Inferring mass in complex scenes by mental simulation. *Cognition*, 157:61–76, 2016.

[10] Tomer D. Ullman, Andreas Stuhlmüller, Noah D. Goodman, and Joshua B. Tenenbaum. Learning physical parameters from dynamic scenes. *Cognitive Psychology*, 104:57–82, 2018.

[11] Gary F Marcus and Ernest Davis. How robust are probabilistic models of higher-level cognition? *Psychological science*, 24(12):2351–2360, 2013.

[12] Ethan Ludwin-Peery, Neil R Bramley, Ernest Davis, and Todd M Gureckis. Broken physics: A conjunction-fallacy effect in intuitive physical reasoning. *Psychological Science*, 31(12):1602–1611, 2020.

[13] Falk Lieder and Thomas L Griffiths. Resource-rational analysis: understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, 43, 2020.

[14] Kevin A. Smith, Peter W. Battaglia, and Edward Vul. Different Physical Intuitions Exist Between Tasks, Not Domains. *Computational Brain & Behavior*, 1(2):101–118, 2018.

[15] F Xu and S Carey. Infants' metaphysics: The case of numerical identity. *Cognitive Psychology*, 30(2):111–153, 1996.

[16] Fei Xu. Categories, kinds, and object individuation in infancy. *Building object categories in developmental time*, pages 63–89, 2005.

[17] Jennifer M Zosh and Lisa Feigenson. Memory load affects object individuation in 18-month-old infants. *Journal of Experimental Child Psychology*, 113(3):322–336, 2012.

[18] Kevin Smith, Lingjie Mei, Shunyu Yao, Jiajun Wu, Elizabeth Spelke, Joshua B. Tenenbaum, and Tomer Ullman. Modeling expectation violation in intuitive physics with coarse probabilistic object representations. In *Advances in Neural Information Processing Systems*, pages 8983–8993, 2019.

[19] Melissa M Kibbe. Varieties of visual working memory representation in infancy and beyond. *Current Directions in Psychological Science*, 24(6):433–439, 2015.

[20] Gerald E Schneider. Two visual systems. *Science*, 1969.

[21] Melvyn A Goodale and A David Milner. Separate visual pathways for perception and action. *Trends in neurosciences*, 15(1):20–25, 1992.

[22] Dwight J Kravitz, Kadharbatcha S Saleem, Chris I Baker, and Mortimer Mishkin. A new neural framework for visuospatial processing. *Nature Reviews Neuroscience*, 12(4):217–230, 2011.

[23] Akira Murata, Vittorio Gallese, Giuseppe Luppino, Masakazu Kaseda, and Hideo Sakata. Selectivity for the shape, size, and orientation of objects for grasping in neurons of monkey parietal area aip. *Journal of neurophysiology*, 83(5):2580–2601, 2000.

[24] Anne B Sereno and John HR Maunsell. Shape selectivity in primate lateral intraparietal cortex. *Nature*, 395(6701):500–503, 1998.

[25] Herbert Edelsbrunner, David Kirkpatrick, and Raimund Seidel. On the shape of a set of points in the plane. *IEEE Transactions on information theory*, 29(4):551–559, 1983.

[26] David Luebke, Martin Reddy, Jonathan D Cohen, Amitabh Varshney, Benjamin Watson, and Robert Huebner. *Level of detail for 3D graphics*. Morgan Kaufmann, 2003.

[27] Albert Michotte. *The Perception of Causality*. Basic Books New York, 1963.

[28] The perception of causality. page 424 p.

[29] David A Rosenbaum. Perception and extrapolation of velocity and acceleration. *Journal of Experimental Psychology: Human Perception and Performance*, 1(4):395, 1975.

[30] JR Tresilian. Perceptual and cognitive processes in time-to-contact estimation: Analysis of prediction-motion and relative judgment tasks. *Perception & Psychophysics*, 57(2):231–245, 1995.

[31] Rob Gray and Ian M Thornton. Exploring the link between time to collision and representational momentum. *Perception*, 30(8):1007–1022, 2001.

[32] Daniel J Simons and Ronald A Rensink. Change blindness: Past, present, and future. *Trends in cognitive sciences*, 9(1):16–20, 2005.

[33] Timothy F Brady, Talia Konkle, Aude Oliva, and George A Alvarez. Detecting changes in real-world objects: The relationship between visual long-term memory and change blindness. *Communicative & integrative biology*, 2(1):1–3, 2009.

[34] Elizabeth S Spelke, Roberta Kestenbaum, Daniel J Simons, and Debra Wein. Spatiotemporal continuity, smoothness of motion and object identity in infancy. *British Journal of Developmental Psychology*, 13(2):113–142, 1995.

[35] Susan M Rivera and Aseen Nancie Zawaydeh. Word comprehension facilitates object individuation in 10-and 11-month-old infants. *Brain research*, 1146:146–157, 2007.

[36] Ilona Bass, Kevin Smith, Elizabeth Bonawitz, and Tomer Ullman. Partial mental simulation explains fallacies in physical reasoning. *PsyArXiv*, 2021.

[37] Edward Vul, Michael C Frank, Joshua B Tenenbaum, and George Angelo Alvarez. Explaining human multiple object tracking as resource-constrained approximate inference in a dynamic probabilistic model. *Advances in neural information processing systems*, 2009.

[38] Herbert Edelsbrunner and Ernst P Mücke. Three-dimensional alpha shapes. *ACM Transactions on Graphics (TOG)*, 13(1):43–72, 1994.

[39] Jonathan F Kominsky, Lewis Baker, Frank C Keil, and Brent Strickland. Causality and continuity close the gaps in event representations. *Memory & Cognition*, 49(3):518–531, 2021.

[40] Hubert D Zimmer. Visual and spatial working memory: from boxes to networks. *Neuroscience & Biobehavioral Reviews*, 32(8):1373–1395, 2008.

[41] Stephen M Kosslyn, William L Thompson, and Giorgio Ganis. *The case for mental imagery*. Oxford University Press, 2006.

[42] Rebecca Keogh and Joel Pearson. The blind mind: No sensory visual imagery in aphantasia. *Cortex*, 105:53–60, 2018.

[43] Christianne Jacobs, Dietrich S Schwarzkopf, and Juha Silvanto. Visual working memory performance in aphantasia. *Cortex*, 105:61–73, 2018.

[44] Rebecca Keogh, Joel Pearson, and Adam Zeman. Aphantasia: The science of visual imagery extremes. In *Handbook of Clinical Neurology*, volume 178, pages 277–296. Elsevier, 2021.

[45] Zoe Pounder, Jane Jacob, Christianne Jacobs, Catherine Loveday, Tony Towell, and Juha Silvanto. Mental rotation performance in aphantasia. *Journal of Vision*, 18(10):1123–1123, 2018.

[46] Matthew J C Crump, John V McDonnell, and Todd M Gureckis. Evaluating Amazon's Mechanical Turk as a Tool for Experimental Behavioral Research. *PLoS ONE*, 8(3):e57410, March 2013.

[47] Eyal Peer, Laura Brandimarte, Sonam Samat, and Alessandro Acquisti. Beyond the turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, 70:153–163, 2017.

[48] Lynn A Cooper. Mental rotation of random two-dimensional shapes. *Cognitive psychology*, 7(1):20–43, 1975.

[49] Jonathan F Kominsky, Brent Strickland, Annie E Wertz, Claudia Elsner, Karen Wynn, and Frank C Keil. Categories and constraints in causal perception. *Psychological Science*, 28(11):1649–1662, 2017.