Behavioral and oculomotor evidence for visual simulation of object movement

Aarit Ahuja

Neuroscience Department, Brown University, Providence, RI, USA

David L. Sheinberg

Neuroscience Department, Brown University, Providence, RI, USA
Carney Institute for Brain Science, Brown University, Providence, RI, USA

We regularly interact with moving objects in our environment. Yet, little is known about how we extrapolate the future movements of visually perceived objects. One possibility is that movements are experienced by a mental visual simulation, allowing one to internally picture an object's upcoming motion trajectory, even as the object itself remains stationary. Here we examined this possibility by asking human participants to make judgments about the future position of a falling ball on an obstacle-filled display. We found that properties of the ball's trajectory were highly predictive of subjects' reaction times and accuracy on the task. We also found that the eye movements subjects made while attempting to ascertain where the ball might fall had significant spatiotemporal overlap with those made while actually perceiving the ball fall. These findings suggest that subjects simulated the ball's trajectory to inform their responses. Finally, we trained a convolutional neural network to see whether this problem could be solved by simple image analysis as opposed to the more intricate simulation strategy we propose. We found that while the network was able to solve our task, the model's output did not effectively or consistently predict human behavior. This implies that subjects employed a different strategy for solving our task, and bolsters the conclusion that they were engaging in visual simulation. The current study thus provides support for visual simulation of motion as a means of understanding complex visual scenes and paves the way for future investigations of this phenomenon at a neural level.

Introduction

We live in a dynamic world that contains an abundance of moving stimuli. To navigate this world, we must interact with these stimuli on a daily basis, and indeed we are able to do so with remarkable sophistication. From reactions to objects already in motion (for example, reaching to catch a football) to acting upon stationary objects to place them in motion (for example, deciding how to roll a bowling ball), we are continuously and often effortlessly carrying out computations of motion. This is especially noteworthy given the extended temporal window inherent to movement. One might then conclude that our ability to interact with moving objects relies crucially upon predictions of what might happen in the future. In spite of how fundamental this ability is to our daily functioning, relatively little is known about how we prospect upon the movements of visually perceived objects. One possible strategy for how one might predict future movement is by internally picturing the likely upcoming trajectory of an object in a visual scene, a faculty we refer to from here on as "visual simulation."

From a theoretical standpoint, one might conceive of visual simulation as being similar to mental imagery, a subject which has a rich history in the field of cognitive neuroscience. As the name suggests, mental imagery refers to our ability to internally envision a known object, even when it is not actually visible to us (Kosslyn, Ganis, & Thompson, 2001). This ability has a tangible biological basis, as numerous studies from the past few decades have shown that early visual areas (such as the primary visual cortex, V1) are active not only when human subjects see visual stimuli, but also when they merely attempt to visualize or imagine these same stimuli with their eyes closed (Kosslyn et al., 2007; Kosslyn, Thompson, Kim, & Alpert, 1995; Kosslyn, Thompson, & Alpert, 1997). Notably, the precise regions of activity across both these conditions (i.e., perception and imagery) overlap in a retinotopically

Citation: Ahuja, A., & Sheinberg, D. L. (2019). Behavioral and oculomotor evidence for visual simulation of object movement. *Journal of Vision*, *19*(6):13, 1–17, https://doi.org/10.1167/19.6.13.

https://doi.org/10.1167/19.6.13 Received January 11, 2019; published June 11, 2019 ISSN 1534-7362

congruent fashion, suggesting that our ability to imagine things has at least some sensory basis. The use of targeted transcranial magnetic stimulation (TMS) over occipital cortex has also been shown to induce deficits in mental imagery, which has further corroborated this idea (Kosslyn et al., 1999). However, most work on mental imagery has relied on the use of relatively static stimuli. A key difference between mental imagery and visual simulation, then, is the essential incorporation of both space and time into internal constructions of the outside world—mental imagery entails generating an internal representation of a static object, whereas visual simulation extends this same idea to dynamic events.

A few previous studies have explored motion imagery and have found that when subjects are shown a moving stimulus and then asked to imagine the same stimulus shortly thereafter, cortical area MT, which is specialized for the perception and processing of motion (Born & Bradley, 2005), is indeed activated (Goebel, Khorram-Sefat, Muckli, Hacker, & Singer, 1998; Emmerling, Zimmerman, Sorger, Frost, & Goebel, 2016). However, given the relatively few studies that have attempted to directly address the question of motion imagery, important questions remain unresolved. For one, both of the previous studies used relatively simple stimuli, as the entire motion trajectory consisted of target displacement in a single, constant direction. As such, whereas the existence of motion imagery and the recruitment of area MT has received some support, how flexible or dynamic it is, and whether it bears specificity is not known. An attempt to decode the imagined motion direction based on fMRI activity in area MT produced surprisingly mixed results, with only two out of fifteen subjects actually exhibiting differentiable activation in an MT ROI (Emmerling et al., 2016). Other work has attempted to explore the question through behavioral demonstrations of the functional consequences of motion imagery (Winawer, Huk, & Boroditsky, 2010; S. Chang & Pearson, 2018). These studies, however, have reached conflicting conclusions, highlighting the need for additional research. It is worth nothing that all of the previous experiments on motion imagery have directed subjects to imagine motion trajectories that were directly cued earlier in the trial (Goebel et al., 1998; Emmerling et al., 2016; S. Chang & Pearson, 2018; Winawer et al., 2010). This raises the possibility that the observed neural correlates may reflect short-term memory retrieval, which is known to reactivate sensory areas involved in motion processing such as MT (Barsalou, 2008; Bisley, Zaksas, Droll, & Pasternak, 2004; Pasternak & Greenlee, 2005). One study did attempt to address this issue by having subjects engage in motion imagery on the basis of predetermined but random rules that dictated how certain exemplars on screen were permitted to move (Kaas, Weigelt, Roebroeck, Kohler, & Muckli, 2010). This study reported that engagement of area MT during this type of motion imagery was surprisingly left lateralized and only observed in half of the recruited subjects (six out of twelve). The reported intersubject variability thus makes it challenging to draw general conclusions. Finally, a number of the aforementioned studies have required subjects to maintain fixation for the entirety of the task (Goebel et al., 1998; Kaas et al., 2010). While this constraint removes confounds directly attributable to oculomotor dynamics, it remains to be seen if and how naturally occurring eye movements interact with imagined movements of objects.

A number of past studies have examined oculomotor dynamics during tasks requiring predictions of motion and future position. For instance, when visually pursuing moving objects, humans are able to make anticipatory eye movements that precede the motion of the pursuit target (Kowler & Steinman, 1979a, 1979b, 1981). Such anticipatory pursuit can be directed either by learning-driven expectation of the target's future motion (Kao & Morrow, 1994), or by symbolic cues in the environment (Kowler, 1989). This suggests that the oculomotor system is able to successfully predict how an object will continue to move in the future and use this information to guide an appropriate motor output. The same has also been shown to be true for saccadic eye movements. When subjects are directed to saccade to a target whose onset is impending but predictable, they initiate their saccades prior to the actual target onset time, thus arriving at the target location with little to no delay (Dallos & Jones, 1963; Stark, Vossius, & Young, 1962). This too demonstrates the oculomotor system's ability to predict and accommodate future events. While such past studies have been illuminating, they also remain subject to certain constraints that limit their generality. First, the stimuli in the aforementioned studies have either been in motion at the time of presentation, or have been presented with a predictable periodicity. In such conditions, the observer has access to informative sensory input that can be used to entrain or initiate predictions of future target position. Whether or not such predictive eye movements might occur with completely static stimuli that lack any motion remains unknown. Further, the anticipatory eye movements reported in these studies have only been shown to "look" a few hundred milliseconds into the future. While this makes sense given the tasks employed, it is unclear whether or not such prospection might extend across a longer temporal window. Visual simulation, as we propose here, incorporates both of these elements (i.e., implementation in the absence of any directed motion cues and prolonged predictions of future movements and events), thus distinguishing our approach from simpler A to B predictions of motion or

future position previously explored in studies on motion imagery and anticipatory eye movements.

Finally, some evidence for our ability to simulate rich and complex motion comes from research on the motor system. For instance, one study showed that a subset of neurons in the macaque premotor cortex (area F5) fired not only when monkeys executed a hand movement, but also when they were made aware of an experimenter executing the same hand movement behind an occluder (Umiltà et al., 2001). The pattern of activity for these neurons in the two conditions was remarkably similar, suggesting that the monkeys were able to extrapolate and simulate a motor plan for a movement that was both extrinsic and not visible to them. A similar study with human participants showed that a readiness potential measured using EEG over the premotor cortex was observed not only when subjects moved their own hand, but also when they observed a video of a moving hand (Kilner, Vargas, Duval, Blakemore, & Sirigu, 2004). Moreover, the onset of this readiness potential actually preceded the hand movement in the video, suggesting that subjects were able to predict the impending motion. This ability has been termed action simulation (Springer, de Hamilton, & Cross, 2012; Springer, Parkinson, & Prinz, 2013). The fact that humans and monkeys are capable of action simulation suggests both species do engage in temporally extended internal constructions of the external world. Importantly, however, research on action simulation has focused largely on the movements of animate actors, and findings of neural correlates have been limited to early motor areas (Kilner et al., 2004; Umiltà et al., 2001; Flanagan & Johansson, 2003; Doerrfeld, Sebanz, & Shiffrar, 2012). Given all that we know about simulation in the motor system, the possible existence of similar, specific and dynamic forms of motion prediction via imagined simulation in the field of vision could be incredibly useful in the improvement of brain-computer interfaces (Banca, Sousa, Duarte, & Castelo-Branco, 2015).

In the present study, we sought to determine whether human subjects might engage in visual simulations of inanimate objects as they take on complicated trajectories of motion spanning one's visual field. We were especially keen on ensuring that all possible attempts at motion imagery and prediction were entirely self-derived, complex, unrestrained, and situated within a naturalistic context. To probe this question, we designed a novel task in which subjects had to make predictions about the future path of a moving ball. As human subjects performed this task, we made behavioral and oculomotor observations. We found that properties of the ball's future motion were good predictors of subjects' behavior, as might be expected if subjects were engaging in a simulation of its motion. We also compared subjects' predictive saccades on a static stimulus to their pursuit of the moving ball. Through this comparison, we found that the eye movements they made while trying to determine the ball's trajectory were remarkably similar to those made while observing the ball execute that same trajectory. We point to these data as likely correlates of visual simulation.

Methods

Participants

Sixteen individuals (seven male, nine female) participated in this study. Participants were compensated a base amount for their time, with additional compensation provided for correct responses on trials. Signed consent was received from all participants. The study was approved by the Brown University IRB.

Task

Each trial began with the presentation of a fixation spot at the center of a blank screen. This was followed by the presentation of a static image (referred to from here on out as a "board"), which comprised one ball at the top, ten semirandomly arranged planks throughout the middle, and two "catchers" at the bottom. The ball and the catchers always appeared in the same position on each board (centered, and just to the left and right of the center, respectively). Figure 1A depicts an example of one such board used in this study (for more information on the pseudorandom board generation procedure, see Supplementary File S1). Upon presentation of this board, participants were asked to judge which of the two catchers the ball would land in, if it were to be dropped from its central position. This required participants to make assessments about how the ball would move as it traversed the field of planks that lay between the ball and the catchers. The physics of our virtual world were determined using Newton Dynamics (http://newtondynamics.com), a "cross-platform life-like physics simulation library." Participants were given no explicit instructions on how to approach this problem, but were told to take as much time as necessary to solve the problem correctly.

Responses were indicated by pressing one of two buttons, each corresponding to one of the two catchers on the screen. Once a response was made, the ball was then dropped, providing participants with visual feedback about their choice. Subjects were instructed to visually pursue the ball as it fell, until it landed in the appropriate catcher. After the ball landed in the catcher, a tone indicated whether the subject had made

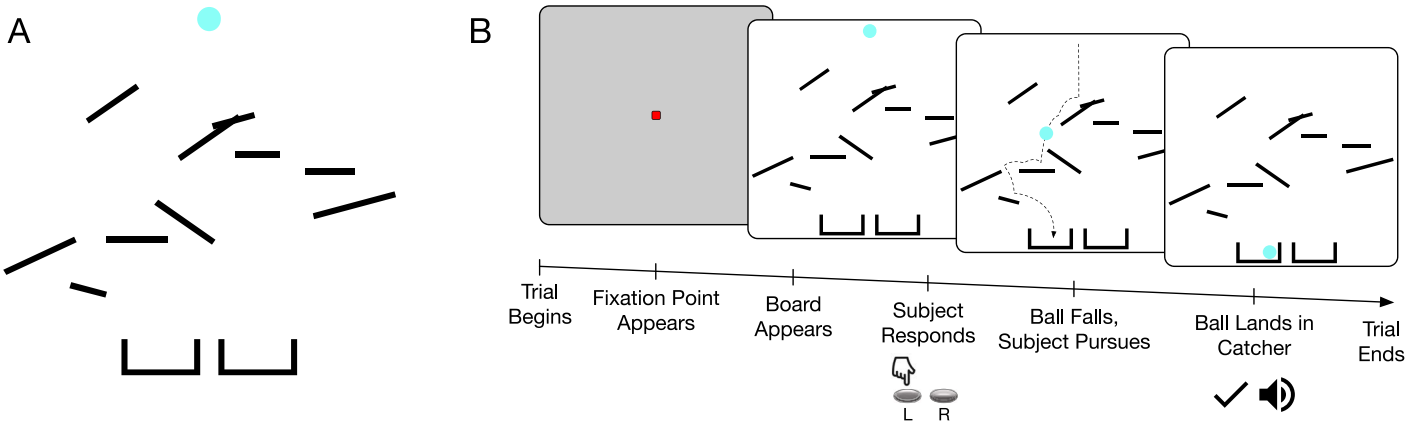

Figure 1. (A) An example the stimuli used. Subjects were shown a static display and asked to judge which catcher the ball would land in, were it to be dropped. (B) An outline of one complete trial.

the correct choice, and the next trial was then initiated with the presentation of the fixation spot. A schematic depicting the progression of a single trial is shown in Figure 1B.

At the start of each session, participants were allowed to practice for as many trials as they desired (generally 10–20) in order to learn the physics of the virtual world being presented, as well as to generally gain familiarity with the progression of a trial. Once subjects reported feeling comfortable in their understanding of the paradigm, we initiated the actual experiment. Each subject was shown 200 unique boards for this experiment. We reused the exact same set of 200 boards (shown in Supplementary Movie S1) for every subject, allowing us to compute board specific metrics averaged across subjects. The proportion of boards on which the ball fell into the left or the right catcher was matched (i.e., 0.5 for each). We also counterbalanced the number of planks the ball interacted with on its trajectory. Specifically, the ball could hit anywhere from one to five planks, resulting in 40 boards per category ($40 \times 5 = 200$). On approximately 2/3 of the boards, the direction of the first change in the ball's trajectory was congruent with the final catcher (i.e., if the initial ball deflection from the midline was towards the left, the answer would be left and vice versa). On these boards, the ball did not cross the midline. On the remaining 1/3, the initial ball deflection was in the opposite direction to the final catcher. On these boards, the ball did cross the midline. Subjects were given a short break halfway through the experiment in order to prevent fatigue. Each session lasted approximately one hour, including both practice and actual trials.

Eye tracking

We used an Eyelink-1000 camera (SR Research) to track participants' eye movements for the entirety of the session. Eye position was sampled at 1 kHz and stored to disk at 200 Hz.

Behavioral analyses

As indicated above, each subject saw exactly the same set of 200 boards. We thus were able to average all sixteen subjects' behavioral data for each board in order to generate a single, high confidence measure of reaction time and accuracy per board. Before doing this, however, we first normalized our data, accounting for the between-subject variability in raw reaction times by transforming each subject's range of reaction times to a 0–100 scale (for more information, see Supplementary File S1). We then averaged these normalized reaction time values in a board-wise fashion. To investigate the question of visual simulation, we asked two key questions for every board: (a) How long of a simulation would be required to mentally recreate the ball's full trajectory for the board? And (b) How much uncertainty would be involved in simulating this trajectory?

To address the first question, we used the number of planks hit by the ball on any trial as a metric for simulation length. This is because as the number of planks hit increases, the total length of the trajectory the ball must travel before arriving at its final destination also increases. Furthermore, each plank hit represents a discrete event that must be factored into the simulation, and thus is likely to contribute to the total length of the simulation process. An example of the relationship between simulation length and the number of planks hit by the ball is shown in Figure 2A. Since we did not place any time constraints on our subjects, we hypothesized that if subjects were engaging in visual simulation, then increasing simulation length would lead to an increase in reaction time, but would have no effect on accuracy.

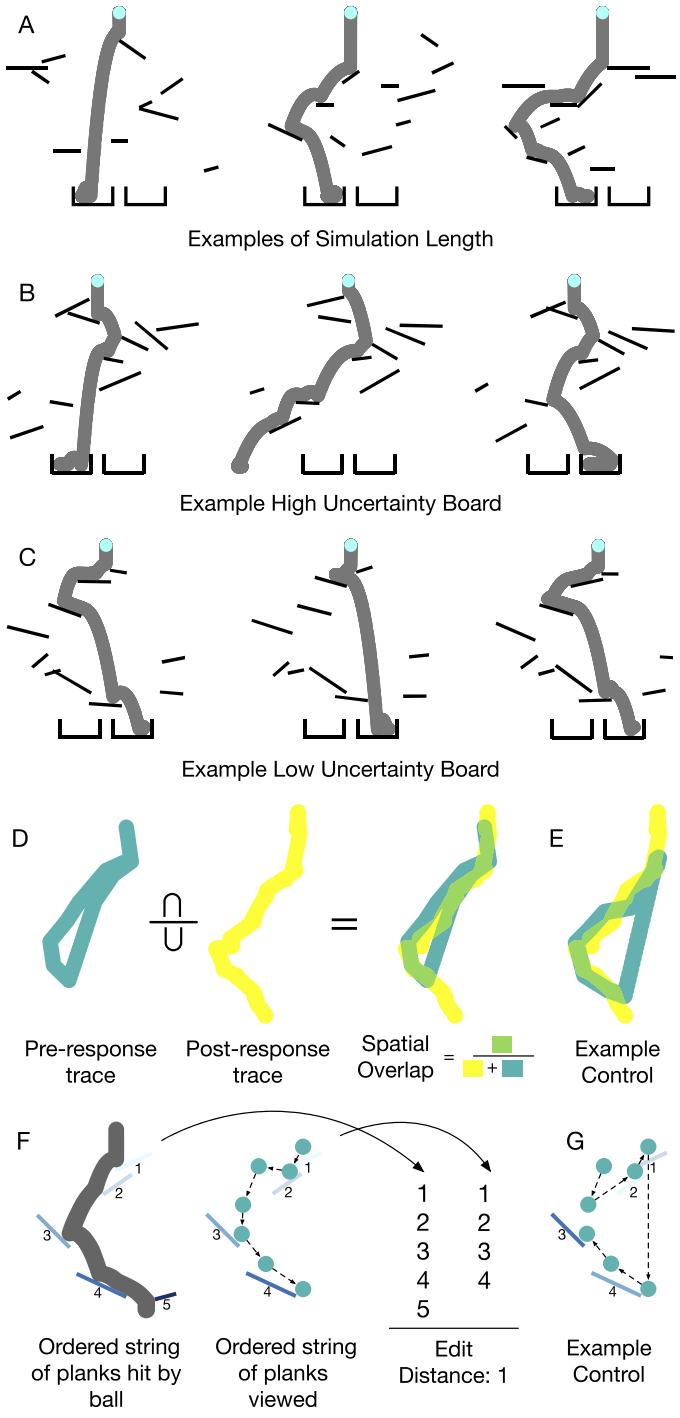

Figure 2. (A) Examples of boards where the ball hit one, three, or five planks (ordered from left to right). The number of planks hit served as an indicator of simulation length, since it dictated both the length of the ball's trajectory, as well as the number of discrete events contained within it. (B) An example of a board where introducing some jitter to the position of the planks had a significant impact on the calculated outcome of the ball's trajectory. Boards like this were assigned high uncertainty scores. (C) An example of a board where this same jitter never changed the calculated outcome of the ball's trajectory. Boards like this were assigned low uncertainty scores. A demonstration of our jitter/uncertainty assignment method can be found in Supplementary Movie S1. (D) A schematic depicting our method for determining spatial overlap. A pre-response saccade trace (left) was overlaid with a post-response smooth pursuit trace (middle), and the intersection of the two was divided by the union (right). This allowed us to assess the degree of spatial similarity between saccades made while determining the ball's final location, and pursuit of the falling ball. (E) As a control, we repeated this same analysis with randomly shuffled, unrelated sets of eye movements to determine a chance level of spatial overlap. (F) A schematic depicting our method for determining temporal overlap. We used edit distance to calculate the sequence similarity between the ordered list of planks hit by the ball and the ordered list of planks looked at by the subjects. (G) As a control, we randomly shuffled the order of the saccades (as indicated by the dotted arrows) to generate a new plank viewing sequence that had no cohesive temporal progression. We then repeated the edit distance calculation with this string to determine a chance level of temporal overlap.

To address the question of simulation uncertainty, we attempted to characterize the number of realistic alternative trajectories one might consider while simulating the ball's path for a particular board. This was based on the assumption that as the number of possible alternatives increased, so would uncertainty. We modeled this uncertainty by introducing some positional jitter to the planks on a given board, and then recalculating the path of the ball on the jittered configuration using our physics engine. This process was carried out offline. Examples of this are shown in Figures 2B and 2C. For some boards (Figure 2B), slight jitter of the planks caused major deviations to the ball's calculated path (relative to the path in the original plank configuration). On others (Figure 2C), jitter had only a minor effect on the trajectory—the ball generally ended up in the same place. We thus used this property of our boards to compute a metric for simulation uncertainty. Specifically, we jittered and recalculated the ball's path for each board 500 times, and then used the proportion of jittered configurations leading to a different outcome (relative to the original configuration) as a metric of simulation uncertainty (for a demonstration of this process, please see Supplementary Movie S2). Boards on which the outcome was rarely altered by plank jitter were classified as being low-uncertainty boards. Conversely, boards on which the outcome was frequently altered by plank jitter were classified as high-uncertainty boards. After assigning each board an uncertainty score, we finally transformed this metric to a 0–100 scale to match our reaction time data. We hypothesized that if subjects were engaging in visual simulation, then they would exhibit higher

reaction times and lower accuracy on high-uncertainty boards relative to low-uncertainty boards.

Oculomotor analyses

In keeping with the paradigm established by past studies on mental imagery (Kosslyn et al., 1995; Kosslyn et al., 1997; Klein et al., 2004; Kosslyn et al., 1999), we compared subjects' pre-response eye movements (i.e., saccades made while viewing the static image of the board) to their post-response eye movements (i.e., smooth pursuit of the ball falling). The basic idea behind this was simply that if the eye movements made while attempting to ascertain the ball's final position showed significant overlap with the eye movements made while perceiving the ball's actual falling trajectory, then one might conclude that subjects visually simulated this movement path in order to solve the task. Note that the eye movements in the pre-response period occur during the static presentation of the board whereas the post-response eye movements occur while the ball smoothly falls toward the catcher. From an oculomotor perspective these are very different, because only saccades will be present in the former, whereas a mixture of pursuit and saccades are likely to occur in the latter. As saccadic and smooth pursuit eye movements have inherently distinct kinematic characteristics (saccades are ballistic and punctuated, whereas smooth pursuit movements are continuous and dependent on the motion attributes of the target), we were unable to rely on traditional measures of oculomotor features such as timing, position, and velocity for our desired comparison. We thus devised two new means of quantifying overlap between pre and post-response eye movements, which were broken down into distinct spatial and temporal domains.

We quantified spatial overlap by overlaying the eye movement traces from the pre-response (hypothesized simulation) and post-response (perception) epochs on top of one another, and then determining the ratio of the intersection and the union of their areas. An example demonstrating this analysis is shown in Figure 2D. We refer to the resulting metric of spatial overlap as simply the intersection, for short. We carried this process out for each trial, and then averaged the resulting intersection values in a subject-wise fashion. This yielded a mean intersection value for each subject. It is important to note that our stimuli do have an inherent directionality to them (top to bottom, left to right, depending on the final position of the ball), which is likely to result in some incidental spatial overlap even for eye movement traces that are entirely unrelated to one another. This is shown in Figure 2E, where we have overlaid two eye movement traces coming from completely separate trials. We were able to capitalize on this form of incidental spatial overlap to quantify a chance intersection level. We did this by simply randomly shuffling the post-response eye movements across trials and redoing the aforementioned intersection analysis on mismatched pairs of traces. In order to ensure that this method of determining chance was sufficiently stringent, we only shuffled traces amongst trials that were matched both in the ball's final position (left vs. right), as well as the number of planks hit by the ball (1–5). We implemented this shuffling protocol for every subject 20 times, and averaged the resulting incidental intersection values on each trial for each iteration. Subsequently, we ended up with a distribution of 20 chance intersection values per subject. We could then compare this distribution to the actual, observed intersection level. We hypothesized that if subjects were engaging in visual simulation, then the degree of spatial overlap between pre-response and post-response eye movements would be greater than that of chance.

To incorporate a temporal dimension in our analysis, we compared pre-response eye movements to the actual trajectory of the ball using a measure known as edit distance. Edit distance indicates the degree of similarity between two alphanumeric strings, with a lower edit distance value reflecting greater similarity (for more details on edit distance, see Supplementary File S1). Thus, if it were possible to discretize subjects' pre-response eye movements into an ordered sequence of numbers, one could use edit distance to compare this to a second string of numbers that reflected the ball's progression in time. An edit distance calculation is particularly well suited for this type of comparison because of it is highly sensitive to order—comparing two strings that contain the same digits in unrelated sequences will result in a high edit distance value, indicating a low degree of similarity. In this context, we used subjects' pre-response eye movements to determine which planks they looked at on the screen, and in what order they did so—this comprised the first string. The ordered list of planks that the ball hit on its trajectory comprised the second string. We then used edit distance to compute the degree of similarity between these two ordered lists on every single trial. This allowed us to probe whether participants were looking at the same planks that the ball was bound to hit, and more importantly, whether the temporal progression of viewing these planks reflected the order in which the ball hit them. This process was carried out for every trial, and the resulting edit distance scores were averaged across trials for each subject. A schematic depicting this procedure is shown in Figure 2F. Since we were primarily concerned with the temporal order of eye movements, we defined chance for every trial by randomly reordering the same eye

movement/plank assignment strings, and then recalculating the edit distance (Figure 2G). This provided a benchmark for the degree of temporal similarity to the ball's path that one might expect even if the eye movements were made in no specific order. This process was repeated 20 times for every trial and the resulting chance edit distance values were averaged in a subject-wise fashion. We thus ended up with a distribution of 20 chance edit distance values per subject. We could then compare this distribution to the actual, observed edit distance value. We hypothesized that if subjects were engaging in visual simulation, then the degree of temporal overlap between pre-response and post-response eye movements would be greater than that of chance.

Finally, we reanalyzed our metrics of spatial and temporal overlap, broken down by whether subjects correctly or incorrectly judged the trial outcome. We expected that if subjects were employing visual simulation to solve this task, then incorrect decisions might be the result of simulating an incorrect path for the ball. We thus hypothesized that the degree of both spatial and temporal similarity between pre-response and post-response eye movements would be lower on incorrect trials relative to correct trials.

Computational analyses

We wondered whether there might be a realistic, nonsimulation based strategy that one could use to solve the present task. An example of such a strategy might involve simply scanning the display to gain a general sense of the overall position and tilt of the intervening planks, and using this heuristic to guide one's answer. In this scenario, one would not have to simulate a moving ball in order to get to the answer, but could instead glean all the information needed via statistical learning of informative, nonsimulation related features of the display. This could subsequently be used to ascertain the appropriate classification via a direct stimulus-response mapping. To see whether such a strategy might indeed be possible, we used a Convolutional Neural Network (CNN). CNNs share numerous organizational motifs with the human visual system, and are known for their exceptional ability to classify images into prelearned categories based purely on salient visual patterns (Fukushima, 1980; LeCun et al., 2008; Ciresan, Meier, & Schmidhuber, 2012). They do this by applying a series of transformations to groups of pixels in an image. In most cases, the end result of these transformations allows the network to ascertain relevant features, and subsequently make the correct classification (Rawat & Wang, 2017). In spite of how successful CNNs often are, subtle manipulations of image properties have also been known to grossly throw off a network's ability to make correct predictions (Szegedy et al., 2013; Nguyen, Yosinski, & Clune, 2014). Human observers, on the other hand, are not affected by these same manipulations, likely owing to various top-down influences (Szegedy et al., 2013). For this reason, it is believed that CNNs are not modeling any real cognitive process, and that they rely purely on the visual elements in a scene. Further, while it is possible to teach a CNN to predict physical dynamics, doing so requires the network to be trained with explicit physical information as opposed to simple images (M. Chang, Ullman, Torralba, & Tenenbaum, 2016; Ehrhardt, Monszpart, Vedaldi, & Mitra, 2017). Given all this, we felt that a CNN could serve as a useful and appropriate tool for emulating an alternate, nonsimulation based strategy like the one described above. For more details on the CNN we employed (including the exact model architecture), please see Supplementary File S1.

We trained the CNN by providing it with images of 75,000 sample boards generated using the same procedure we used to create the 200 boards used in the actual task. For each board image, we assigned the correct response based on the physical simulation. From these training data, the network generated a model, which we then evaluated using the same set of 200 boards that had been shown to our human subjects. Strikingly, we found that the CNN's model was able to predict the correct answer for the 200 board set with 84% accuracy. This falls comfortably within the range of accuracy values we observed with our human subjects (as shown in Figure 5A). Thus having confirmed that an alternate strategy did in fact exist, we wanted to see if this strategy too might predict our subjects' behavioral data. Our model provided its prediction output in the form of two probabilities, indicating the likelihood that the answer was either left, $P_{(L)}$, or right, $P_{(R)}$. Since there were only two possible options, the sum of these probabilities was always 1. We were able to use these probabilities to devise a new way of assigning an uncertainty score to each board, within the context of this particular strategy. We determined this alternate uncertainty score with the following formula: $Uncertainty = 1 - |P_{(L)} - P_{(R)}|$. Thus, a board for which the model predicted a $P_{(L)}$ value of 0.99 and $P_{(R)}$ value of 0.01 would be classified as being low uncertainty. On the other hand, a board for which the model predicted a $P_{(L)}$ value of 0.51 and a $P_{(R)}$ value of 0.49 would be classified as being high uncertainty. Here again, we transformed our uncertainty scores to a 0–100 scale. We then analyzed whether this new, image-analysis based method of assigning board uncertainty was predictive of reaction time and accuracy, and if so, how it compared to the simulation-based method of assigning uncertainty described above.

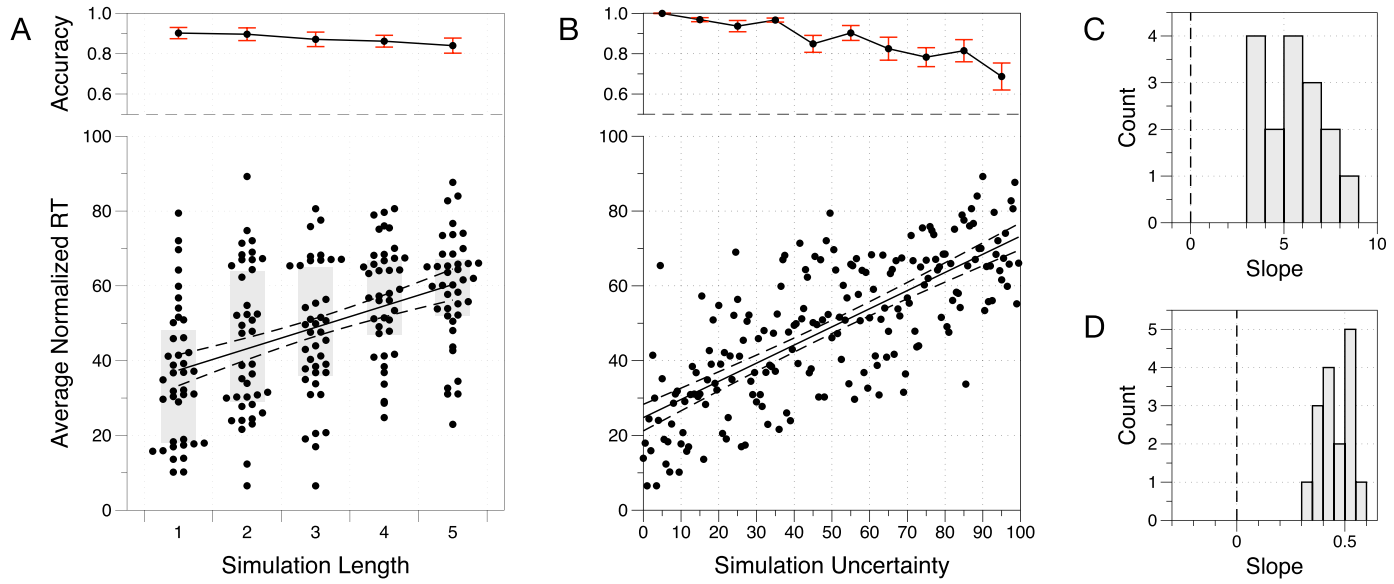

Figure 3. (A) Across subjects average normalized reaction time and accuracy as a function of the simulation length value assigned to every board. Each black point represents one board from the set of 200 boards that was shown to all subjects. The gray shaded regions represent the 1st–3rd quartile of each distribution. (B) Across subjects average normalized reaction time and accuracy as a function of the simulation-based uncertainty value assigned to every board. In both (A) and (B), the solid line represents the slope of the linear regression, and the dotted lines represent the 95% confidence interval (CI) for the slope of the regression line. The red bars in the accuracy sections of both graphs represent the standard error of the across-subject means for the boards falling in each category/bin. (C) A histogram showing the slopes of the regression in (A) when carried out with each individual subject's data instead of sample wide averages. (D) A histogram showing the slopes of the regression in (B) when carried out with each individual subject's data instead of sample wide averages.

Results

Behavioral evidence of visual simulation

To assess whether subjects were employing visual simulation in this task, we classified each board based on the potential length of and uncertainty associated with the simulation that would have to be carried out on that board. We then compared these board characteristics to average normalized figures of reaction time and accuracy. Figure 3A depicts the effect of simulation length on these behavioral parameters. Using a simple linear regression, we found that simulation length predicted reaction time on this task, $F(1, 198) = 44.39$, $p < 0.001$, $R^2 = 0.1831$. This is congruent with our hypothesis, and is compatible with the idea that subjects were indeed carrying out visual simulations. To ensure that this effect was not being driven primarily by a small subset of subjects (note that the previous regression was carried out using the mean reaction times of all of our subjects), we repeated this same analysis on a subject-by-subject basis. We found that simulation length was a significant predictor of reaction time for each individual subject. The distribution of the slopes for the 16 individual regressions is shown in Figure 3C. A single sample t test revealed that the mean of this distribution was significantly greater than zero, $t(15) = 15.28$, $p < 0.001$. We also noted that accuracy on this task was not significantly predicted by simulation length, $F(1, 198) = 2.406$, $p = 0.122$, $R^2 = 0.012$. This too is unsurprising, since subjects' overall accuracy was very high. Further, as we did not impose any time constraints on our subjects, the effect of the speed-accuracy trade-off was largely reflected in the reaction time, with no notable effect on accuracy. Figure 3B depicts the effect of simulation uncertainty on reaction time and accuracy. Here we note that as simulation uncertainty increased, so did reaction time $F(1, 198) = 240.5$, $p < 0.001$, $R^2 = 0.5485$). As before, we repeated this same analysis on a subject-by-subject basis. We found that simulation uncertainty was a significant predictor of reaction time for each individual subject. The distribution of the slopes for the 16 individual regressions is shown in Figure 3D. A single sample t test revealed that the mean of this distribution was significantly greater than zero, $t(15) = 24.258$, $p < 0.001$. Finally, an increase in simulation uncertainty predicted a decrease in accuracy, $F(1, 198) = 44.46$, $p < 0.001$, $R^2 = 0.1834$. These findings too are consistent with our hypotheses.

Because simulation length was not a clear predictor of accuracy on this task while simulation uncertainty was,

we were interested in a possible interaction between the two measures that might have been missed in the linear regression models. We thus implemented a multiple regression model of simulation length, uncertainty, and their interaction onto subjects' accuracy. We found that while simulation uncertainty was a significant predictor of accuracy ($\beta = -0.0050111$, $p < 0.001$), this was not the case for length ($\beta = 0.0027728$, $p < 0.889$), or the interaction term ($\beta = 0.0004302$, $p < 0.224$). The results of this multiple regression model confirm that simulation length and uncertainty do indeed have differential effects on task accuracy. In summary, we found that mean reaction time and accuracy across subjects on this task could successfully be predicted by metrics describing the ball's trajectory. Our data thus support the idea that subjects were carrying out visual simulations as a strategy for solving this task.

Oculomotor evidence of visual simulation

Our oculomotor analyses were based on the hypothesis that if subjects were carrying out visual simulations, then their eye movements in the pre-response period (when attempting to ascertain the final position of the ball) ought to bear a high degree of similarity to their eye movements in the post-response period (when actually perceiving and pursuing the falling ball). We split this comparison into two distinct dimensions—spatial and temporal. Figure 4A depicts the percentage of observed spatial overlap between pre and post-response eye movements relative to a chance distribution of values, broken down in a subject-wise manner (for details on how each subject's chance distribution was generated, see "Oculomotor analyses" section of Methods). As is clear from this figure, the majority of subjects showed a far greater degree of spatial overlap between pre- and post-response eye movements than would be expected by chance. This result indicates that eye movements made while prospecting upon the ball's movement greatly resembled those made while perceiving the ball's movement and suggests that subjects were actively simulating the ball's future path in the pre-response period. Figure 4B replots this same data, but with the boxplots in Figure 4A averaged to a single chance value (represented by the black points) for each subject. A paired samples t test of the means of the intersection in the actual and shuffled conditions revealed a significant difference between the two, $t(15) = 6.6626$, $p < 0.001$.

Figure 4C depicts the degree of similarity in the temporal progression of pre-response eye movements and the ball's trajectory relative to chance, broken down in a subject-wise manner (for details on how each subject's chance distribution was generated, see "Oculomotor analyses," above). Recall that for the edit distance metric used here, a lower value indicates greater similarity. We found that almost all subjects showed a significantly greater degree of temporal overlap than would be expected by chance. This is congruent with our analyses of spatial overlap, and provides further evidence suggesting that subjects were likely carrying out visual simulations. Figure 4D shows this same data, but with the boxplots in Figure 4C averaged to a single chance value (represented by the black points) for each subject. A paired t test of the means of the edit distance in the actual and shuffled conditions revealed a significant difference between the two, $t(15) = -7.7454$, $p < 0.001$.

Figure 4E and 4F show the same metrics of spatial and temporal similarity, but with trials sorted by correct or incorrect responses. We found that both spatial, $t(15) = 2.1525$, $p < 0.05$, and temporal, $t(15) = -3.7013$, $p < 0.001$, similarity between pre- and post-response eye movements was lower on trials that subjects incorrectly responded to compared to correct trials. This finding is in line with our hypothesis, and suggests that one possible factor explaining why subjects may have incorrectly responded on a trial is that they simulated the wrong ball path, leading to the wrong answer. Finally, we separately analyzed and compared each subject's degree of spatial and temporal overlap across the first 20 and last 20 trials of the session. This comparison is crucial because subjects were explicitly asked to pursue the falling ball, raising the possibility that over time, this instruction might have implicitly trained them to use eye movements to predict where the ball would fall. A paired t test of the mean intersection of the first 20 versus last 20 trials for each subject showed no significant difference, $t(15) = -0.89119$, $p = 0.3869$. A paired t test of the mean edit distance of the first 20 versus last 20 trials for each subject yielded the same outcome, $t(15) = 0.71735$, $p = 0.4842$. Based on these results, we do not consider it likely that repeatedly pursuing the falling ball necessarily led to an evolution in strategy or entrainment of simulation.

Computational evidence of visual simulation

To investigate possible alternate strategies, we trained a convolutional neural network on this task. The network's model was able to predict the correct answer with 84% accuracy on the same boards that we showed to our human subjects, indicating that at least one alternate, neurally plausible, nonsimulation-based strategy does exist. Using our CNN model outputs, we computed a new measure of uncertainty for each of the 200 boards (for more detail, please see "Computational analyses" section of Methods). We then asked if the CNN-derived uncertainty metric predicted human

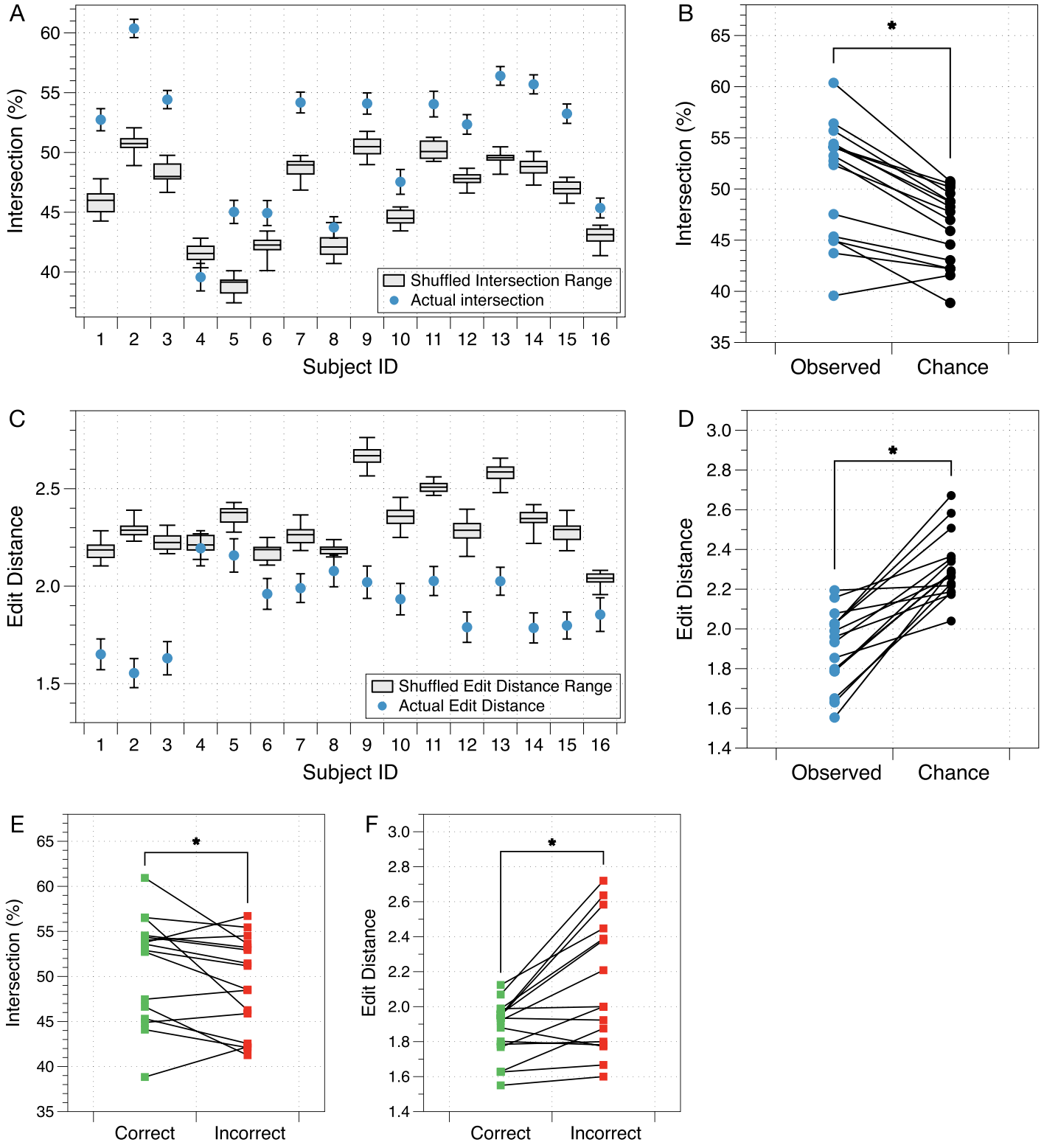

Figure 4. (A) A breakdown of each subject's chance intersection values versus their actual intersection value. Box plots represent a distribution of twenty chance intersection values generated by shuffling (whiskers span maximum to minimum), and blue points represent actual mean intersection values. (B) Pairwise comparisons of chance versus actual intersection values for each subject. Black points represent the average of each box plot in (A). (C) and (D) Same as (A) and (B), but for edit distance instead of intersection. (E) and (F) Pairwise comparisons of intersection and edit distance values on trials that subjects got correct versus incorrect.

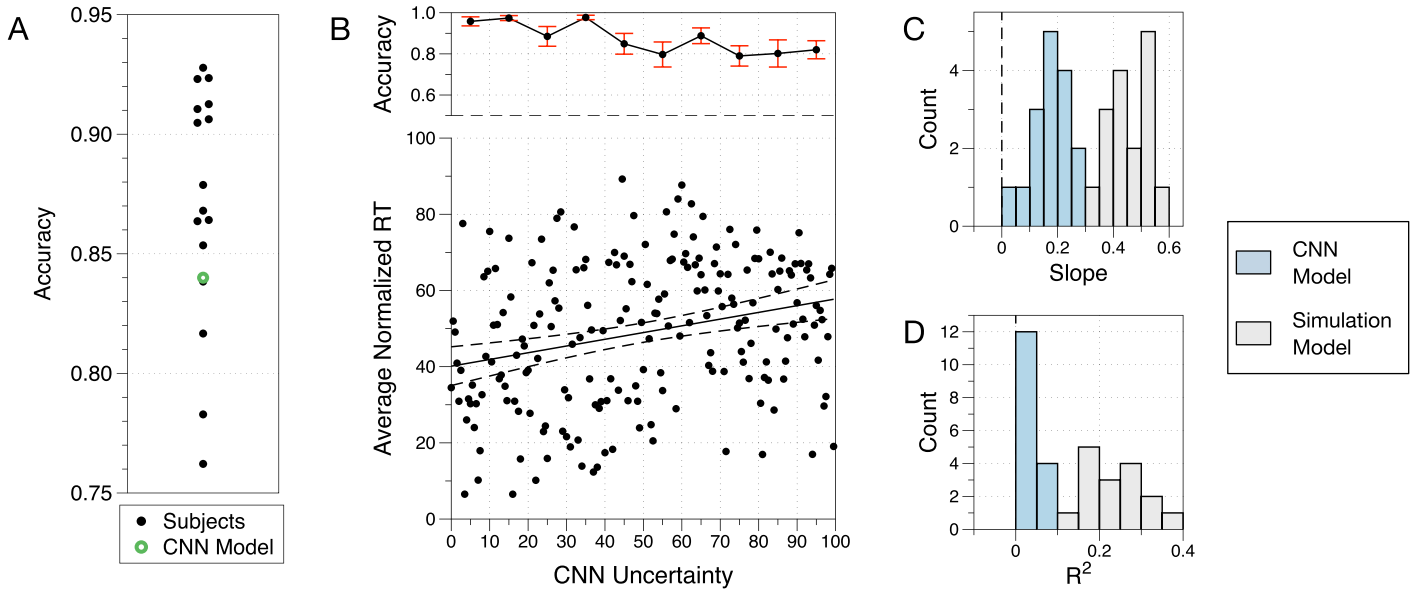

Figure 5. (A) Subjects' accuracy on this task versus the CNN model's accuracy. (B) Across subjects average normalized reaction time and accuracy for each board as a function of the CNN-based uncertainty value assigned to that board. The dotted lines represent the 95% CI for the slope of the regression line. (C) A histogram showing the slopes of the regression of reaction time onto CNN uncertainty (as shown in 5B) and simulation uncertainty (as shown in 3B) when carried out with each individual subject's data instead of sample wide averages. (D) A histogram showing the same comparison as in (C), but with the R^2 values for each model.

behavior, and if so, how this relationship compared to the previously shown relationship between behavior and simulation-derived uncertainty. Figure 5B depicts the same across-subject averaged behavioral data as in Figure 3B, now plotted as a function of CNN uncertainty as opposed to simulation uncertainty. We found that CNN uncertainty was in fact predictive of subjects' average reaction times, $F(1, 198) = 15.49$, $p < 0.001$, $R^2 = 0.072$, and accuracy, $F(1, 198) = 16.72$, $p < 0.001$, $R^2 = 0.077$, on this task. We thus had two possible, valid models for explaining subject behavior—one based on simulation of the ball's trajectory, and the other based purely on computations of the spatial relationships between onscreen objects. To distinguish between these two possibilities, we assessed how much variance in subject behavior was accounted for by the two uncertainty metrics pertaining to each strategy (i.e., simulation uncertainty in Figure 3B and CNN uncertainty in Figure 5B. We note that the variance in reaction times explained by the CNN model is extremely small ($R^2 = 0.072$), whereas the variance explained by the simulation model is far greater ($R^2 = 0.5485$). The same applies for task accuracy (i.e., R^2 values of 0.07 and 0.18 for the CNN and simulation models respectively). The fact that the simulation model is a much better predictor of subjects' behavior strongly suggests that subjects were likely engaging in visual simulation. To corroborate this finding, we ran these same regression analyses on a subject-by-subject basis. We found that CNN uncertainty was in fact not a significant predictor of reaction time for six of our subjects, whereas simulation uncertainty was a significant predictor for all 16 subjects. A comparison of the slopes and R^2 values of these two regression models across subjects shows that the simulation uncertainty model consistently yielded significantly higher slope, $t(15) = 16.931$, $p < 0.001$, and R^2 values, $t(15) = 12.31$, $p < 0.001$. Notably, for the majority of subjects (12 out of 16), the R^2 value for the CNN model barely exceeded 0. The nonoverlapping distributions of these values for all 16 subjects is shown in Figure 5C and 5D. Finally, calculating the Akaike information criterion for both models returned a lower value for the simulation model compared to the CNN model (Δ AIC: 143.956), further supporting the idea that subjects were likely simulating the ball's motion trajectory. Overall, we conclude that while there may be various valid approaches to solving this task, our subjects' behavior is best explained by a visual simulation strategy as opposed to a global image analysis strategy that might be exploited by a CNN.

Discussion

In the present study, we were interested in obtaining evidence for visual simulation—a process through which one might be able to internally imagine the upcoming motion trajectory of an object in her or his visual field. We liken this ability to mental imagery, except with a dynamic internal representation of the external word, as opposed to a static one. The result of

this simulation process (i.e., an internal imagery-like representation of a complex motion) may then serve as a guide that one can rely on to predict an upcoming trajectory and direct an appropriate behavioral response. We also posit that a correlate of this cognitive phenomenon exists in the eye movements made while one engages in a simulation. We thus relate the internal, imagery-based aspect of the simulation process to an explicit, observable motor output. We tested these ideas by designing a novel task in which subjects had to ascertain the path of a ball that was subject to the normal laws of gravity. This task could, in theory, be solved by employing visual simulation. As subjects performed this task, we recorded their eye movements as well as behavioral metrics such as reaction time and accuracy. We found that subjects' eye movements made while determining the ball's final location overlapped heavily with their eye movements made while perceiving the same trajectory just a few seconds later. We also found that reaction time and accuracy on our task were predicted by the properties of the ball's trajectory. Together, these findings indicate subjects were engaging in a temporally extended construction of an imagined motion (observable through a systematic series of eye movements), which they then used to inform their responses. This provides overt evidence that humans are in fact capable of employing visual simulation to solve problems of complex motion prediction. Of course, this does not mean that visual simulation is the only strategy people employ to predict motion in one's daily life, especially because predictions of motion occur at various levels of abstraction, and at various levels of spatial and temporal resolution. We suggest that this ability may be most likely employed when prospecting upon the movements of objects that are at rest but may move in the future, especially in time-permitting contexts (for example, when deciding how to navigate carrying a couch up a winding staircase).

It is worth noting that we designed our task to obey the laws of basic Newtonian physics. We are certainly not the first to adopt this strategy—indeed a number of previous studies have specifically investigated how our intuitive understanding of physics can be used to solve cognitive problems (Battaglia, Hamrick, & Tenenbaum, 2013; Fischer, Mikhael, Tenenbaum, & Kanwisher, 2016; Ullman, Spelke, Battaglia, & Tenenbaum, 2017). In most of these studies, however, the motion of the stimulus has been relatively simple, and the key cognitive process in focus has been physical reasoning. Our motivation in the present study differed from the previous studies in that we were primarily interested in the process of predicting a complex motion, and we relied on people's understanding of rigid body dynamics simply to arrive upon an easily understood, common set of rules governing this motion. Further, previous research on predictive pursuit has shown that subjects are able to make more anticipatory responses when the spatial cues indicating an upcoming motion are based in physical rules as opposed to arbitrary ones (Kowler, Aitkin, Ross, Santos, & Zhao, 2014). This research has focused on smooth pursuit of already moving stimuli, whereas we have extended this finding to saccades on a static display. While the physical basis of our approach does place some limitations on the breadth of conclusions we are able to draw, we felt it was nonetheless useful, especially in the early stages of investigating visual simulation. In future experiments we intend to broaden the scope of motion types to better understand whether visual simulation remains viable when motion properties are not congruent with physics.

A significant line of evidence we present in support of visual simulation comes from our oculomotor analyses. Past studies on mental imagery and action simulation have long relied on comparing neural correlates across two conditions, one involving visual perception, and the other involving the cognitive phenomenon of interest (Kosslyn et al., 1995; Kosslyn et al., 1997; Klein et al., 2004; Kosslyn et al., 1999). In the present study, we adopted this same idea, using oculomotor measures in place of neural recordings. The concept of utilizing eye movements as a window into cognition is not new, and studies on scan paths during complex tasks date back over four decades (Noton & Stark, 1971; Chase & Simon, 1973). In fact, the theory of deictic coding posits that eye movements can serve to orient a cognitive process by anchoring it in the physical world (Ballard, Hayhoe, Pook, & Rao, 1997). Past research on action simulation has also shown that when tracking an actor carrying out a series of hand movements, subjects make predictive saccades as opposed to reactive ones, indicating an ability to orient gaze to accommodate future events (Flanagan & Johansson, 2003). Similarly, other work examining the pursuit of already moving objects has shown that people are capable of making predictive saccades based on where the object of interest is likely to be in the near future (Diaz, Cooper, Rothkopf, & Hayhoe, 2013). With this in mind, it then makes complete sense that the process of carrying out a visual simulation of an object's future motion trajectory (that spans significant distance in visual space) could be read out through a sequence of eye movements. Here, we have shown that there is a remarkable degree of similarity in the spatial organization of eye movements made while attempting to determine a ball's future motion and those made while perceiving that motion. Further, we have demonstrated that the temporal order of these eye movements bears a degree of overlap with the progression of the ball's path that far exceeds chance. The improvement in temporal granularity gained by looking at sequences of saccades (relative to simple

metrics such as reaction time) is especially important, since it allows one to tie discrete events in the pre-response period to discrete properties of the upcoming motion. Our emphasis on understanding the properties of unrestrained eye movements during the simulation process distinguishes this research from past experiments on motion imagery where subjects have been required to maintain fixation (Goebel et al., 1998; Kaas et al., 2010). The predictive eye movements we have shown here also differ fundamentally from past work on anticipatory saccades and pursuit in that (a) they were entirely internally derived (i.e., there was no existing sensory motion to entrain to), and (b) they were carried out across an extended window of time (on the order of many seconds). Altogether, we conclude that eye movements appear to be indicative of a systematic attempt to plot out a future motion, as one might be expected to do when engaging in visual simulation.

Of course, we are also able to look at broader behavioral metrics (in this case reaction time and accuracy) as a function of the stimulus' properties. It is worth pointing out that while simulation length was predictive of reaction time, it did not predict accuracy. However, this is not entirely surprising, since subjects were not given a time limit within which they had to respond, and so any effect of a speed-accuracy trade-off would be reflected primarily in reaction times. Further, it is not guaranteed that an increase in simulation length would actually lead to an increase in perceived difficulty on our task. To provide an analogy, if asked to count to 10 or to 100, one would take longer to complete the latter because the process has more steps, but that doesn't necessarily mean that more errors would be made as a result of the longer process. Even if the probability of making counting errors were greater in the case of counting to 100, these errors would likely not be frequent enough to result in a statistically significant decrement (since people are generally quite good at counting). This is in fact exactly what we observed—overall, people were very good at our task, and although there was a mild negative trend of decreasing accuracy with increasing simulation length, this trend was not statistically significant.

On the other hand, simulation uncertainty predicted both reaction time and accuracy. Here the decrease in accuracy is expected (even though there were no time constraints) because our metric of uncertainty was primarily driven by the number of plausible alternatives on a given board. As the number of viable possibilities increased, the likelihood of ultimately picking the correct one was reduced. Overall, the fact that the properties of the ball's trajectory could be used to predict reaction time and accuracy is noteworthy because it demonstrates a direct relationship between subjects' behavior and the complex future movements that they had to ascertain, which is precisely what one might expect if people were carrying out visual simulation.

Our final line of analyses was aimed at addressing two questions: (a) are there any potential alternative strategies one might adopt to solve this task, and if so, (b) might they be used to explain our behavioral data. We utilized a convolutional neural network to emulate a possible alternative strategy because such networks are organizationally similar to the human visual system, but lack any explicit ability to carry out physics-driven simulations. Additionally, past work with neural networks has shown a striking congruence with both human behavior and neurophysiological data obtained from areas within the visual pathway, making them a powerful computational modeling tool within the field of neuroscience (Yamins et al., 2014). We found that our very simple—off the shelf—network was able to successfully predict correct outcomes on the set of 200 boards at a level consistent with that of our human participants. This allowed us to answer the first question and conclude that there is at least one other possible strategy for tackling our task. The limitation to this type of analysis is that we only have a rough understanding of the operations occurring within the hidden layers of the network, and research is still ongoing into the question of how or why CNNs perform as well as they do (Yosinski, Clune, Nguyen, Fuchs, & Lipson, 2015). We are thus unable to state with complete certainty how exactly the network solved this task. However, given what we do know about the basic structure and function of CNNs, it seems likely that the network was able to ascertain informative features of the overall plank configuration on every board, and use this information to generate probabilities for the two possible answers (i.e., left and right). We hope to probe this question further in future iterations of this study to pinpoint what exactly these features might be, and whether they can be related to salience in perception among human observers.

To address our second question, we utilized the model's output metrics to devise a new, nonsimulation dependent method of assigning uncertainty values to our boards. We then simply regressed our behavioral data onto this metric to see if it could predict subjects' reaction times and performance on the task. Interestingly, we did find a relationship between our behavioral data and the CNN-based uncertainty metric, although this relationship was extremely weak at the population level and did not achieve significance for two-thirds of our subjects when tested at the individual level. This is in stark contrast to the strong relationship we observed using our simulation-based uncertainty metric using both population and individual subject data. It is not entirely surprising that the CNN model predicted some variance in some subjects' behavior, since the positional

relationships between the planks on screen do indeed affect the ball's actual trajectory. However, we have shown that this low-level framework for extrapolating uncertainty is only loosely related to subject behavior, and that a model that factors in the ball's trajectory as determined through simulation is much better at explaining reaction time and accuracy at this task. From this, we conclude that our subjects were not simply generally scanning plank configuration patterns for relevant clues, but were indeed simulating a dynamic motion trajectory.

This finding also raises the possibility that familiarity plays a role in determining the strategy that people rely on. Previous research with chess players has shown that while less advanced players tend to engage in more time-consuming, "look-ahead" strategies in order to determine their next move, grandmasters are often able to make rapid, high quality moves even when only briefly presented with the configuration of the pieces on the board (Holding & Reynolds, 1982; Calderwood, Klein, & Crandall, 1988; Gobet & Simon, 1996). Since our human subjects were not familiar with the set-up of this task prior to their participation in the experiment, one might conceive of them as relative novices who would need to "look-ahead" or simulate possible outcomes in order to arrive upon an answer. Our network, on the other hand, was trained on 75,000 example boards, making it somewhat analogous to a grandmaster who would have likely witnessed various chess piece configurations many thousands of times. It is fair to hypothesize, then, that with extended practice, human subjects' behavior on this task might be predicted by the network's outputs. It is also entirely possible that there are other strategies we have not yet considered for explaining human behavior on this task. Given the current evidence, however, we conclude that subjects were carrying out visual simulations of the ball's future motion path.

Conclusion

In the present study, we have combined three complementary lines of evidence—behavioral, oculomotor, and computational—to demonstrate that human subjects can and do engage in visual simulation as a strategy for predicting the future motion of objects. A deeper understanding of motion simulations in the visual system has the potential to not only improve our grasp of the brain as a whole, but could also provide valuable enhancements to existing brain-computer interfaces. In future studies, we plan to characterize this phenomenon at a neural level in both humans and nonhuman primates.

Keywords: imagery, visual simulation, eye movements, object motion, motion prediction

Acknowledgments

We would like to acknowledge Shaobo Guan and Dr. Theresa Desrochers for their many intellectual contributions during the analysis phase of this research. We would also like to thank Mark Hedinger and Oscar Machado for their help with running subjects. Finally, we would like to thank Diana Burk, Ruobing Xia, Ryan Miller, Wenhao Dang, John Ghenne, Nadira Yusif-Rodriguez, and Dr. Theresa McKim for all of their various suggestions and insights. This research was supported by National Science Foundation Grant 1632738 to David L. Sheinberg, National Institutes of Health Grant R01EY14681 to David L. Sheinberg, and National Institutes of Health Vision Training Grant 2T32EY018080-11 to Michael A. Paradiso.

Commercial relationships: none.
Corresponding author: David L. Sheinberg.
Email: david_sheinberg@brown.edu.
Address: Neuroscience Department, Brown University, Providence, RI, USA.

References

Ballard, D. H., Hayhoe, M. M., Pook, P. K., & Rao, R. P. (1997). Deictic codes for the embodiment of cognition. *Behavioral and Brain Sciences*, *20*(4), 723–767.

Banca, P., Sousa, T., Duarte, I. C., & Castelo-Branco, M. (2015). Visual motion imagery neurofeedback based on the hMT+/V5 complex: Evidence for a feedback-specific neural circuit involving neocortical and cerebellar regions. *Journal of Neural Engineering*, *12*(6), 066003, https://doi.org/10.1088/1741-2560/12/6/066003.

Barsalou, L. (2008). Grounded cognition. *Annual Review of Psychology*, *59*(1), 617–645, https://doi.org/10.1146/annurev.psych.59.103006.093639.

Battaglia, P., Hamrick, J., & Tenenbaum, J. (2013). Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences, USA*, *110*(45), 18327–18332, https://doi.org/10.1073/pnas.1306572110.

Bisley, J. W., Zaksas, D., Droll, J. A., & Pasternak, T. (2004). Activity of neurons in cortical area MT during a memory for motion task. *Journal of*

Neurophysiology, *91*, 286–300, https://doi.org/10.1152/jn.00870.2003.

Born, R. T., & Bradley, D. C. (2005). Structure and function of visual area MT. *Annual Review of Neuroscience*, *28*(1), 157–189, https://doi.org/10.1146/annurev.neuro.26.041002.131052.

Calderwood, R., Klein, G., & Crandall, B. (1988). Time pressure, skill, and move quality in chess. *The American Journal of Psychology*, *101*(4), 481, https://doi.org/10.2307/1423226.

Chang, M., Ullman, T., Torralba, A., & Tenenbaum, J. (2016). A compositional object-based approach to learning physical dynamics. *arXiv*1612.00341.

Chang, S., & Pearson, J. (2018). The functional effects of prior motion imagery and motion perception. *Cortex*, *105*, 83–96, https://doi.org/10.1016/j.cortex.2017.08.036.

Chase, W., & Simon, H. (1973). Perception in chess. *Cognitive Psychology*, *4*(1), 55–81, https://doi.org/10.1016/0010-0285(73)90004-2.

Ciresan, D., Meier, U., & Schmidhuber, J. (2012). *Multi-column deep neural networks for image classification. 2012 IEEE Conference on Computer Vision and Pattern Recognition* (*Vol. 1*, pp. 3642–3649), https://doi.org/10.1109/CVPR.2012.6248110.

Dallos, P., & Jones, R. (1963). Learning behavior of the eye fixation control system. *IEEE Transactions on Automatic Control*, *8*(3), 218–227, https://doi.org/10.1109/TAC.1963.1105574.

Diaz, G., Cooper, J., Rothkopf, C., & Hayhoe, M. (2013). Saccades to future ball location reveal memory-based prediction in a virtual-reality interception task. *Journal of Vision*, *13*(1):20, 1–14, https://doi.org/10.1167/13.1.20. [PubMed] [Article]

Doerrfeld, A., Sebanz, N., & Shiffrar, M. (2012). Expecting to lift a box together makes the load look lighter. *Psychological Research*, *76*(4), 467–475, https://doi.org/10.1007/s00426-011-0398-4.

Ehrhardt, S., Monszpart, A., Vedaldi, A., & Mitra, N. (2017). Learning to represent mechanics via long-term extrapolation and interpolation. *arXiv*1706.02179.

Emmerling, T., Zimmermann, J., Sorger, B., Frost, M., & Goebel, R. (2016). Decoding the direction of imagined visual motion using 7T ultra-high field fMRI. *NeuroImage*, *125*, 61–73, https://doi.org/10.1016/j.neuroimage.2015.10.022.

Fischer, J., Mikhael, J. G., Tenenbaum, J. B., & Kanwisher, N. (2016). Functional neuroanatomy of intuitive physical inference. *Proceedings of the National Academy of Sciences, USA*, *113*(34), E5072–E5081, https://doi.org/10.1073/pnas.1610344113.

Flanagan, R., & Johansson, R. (2003, August 14). Action plans used in action observation. *Nature*, *424*(6950), 769–771, https://doi.org/10.1038/nature01861.

Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, *36*(4), 193–202, https://doi.org/10.1007/BF00344251.

Gobet, F., & Simon, H. A. (1996). The roles of recognition processes and look-ahead search in time-constrained expert problem solving: Evidence from grand-master-level chess. *Psychological Science*, *7*(1), 52–55, https://doi.org/10.1111/j.1467-9280.1996.tb00666.x.

Goebel, R., Khorram-Sefat, D., Muckli, L., Hacker, H., & Singer, W. (1998). The constructive nature of vision: Direct evidence from functional magnetic resonance imaging studies of apparent motion and motion imagery. *European Journal of Neuroscience*, *10*(5), 1563–1573, https://doi.org/10.1046/j.1460-9568.1998.00181.x.

Holding, D., & Reynolds, R. (1982). Recall or evaluation of chess positions as determinants of chess skill. *Memory & Cognition*, *10*(3), 237–242, https://doi.org/10.3758/BF03197635.

Kaas, A., Weigelt, S., Roebroeck, A., Kohler, A., & Muckli, L. (2010). Imagery of a moving object: The role of occipital cortex and human MT/V5+. *NeuroImage*, *49*(1), 794–804, https://doi.org/10.1016/j.neuroimage.2009.07.05.

Kao, G., & Morrow, M. (1994). The relationship of anticipatory smooth eye movement to smooth pursuit initiation. *Vision Research*, *34*(22), 3027–3036, https://doi.org/10.1016/0042-6989(94)90276-3.

Kilner, J., Vargas, C., Duval, S., Blakemore, S.-J., & Sirigu, A. (2004). Motor activation prior to observation of a predicted movement. *Nature Neuroscience*, *7*(12), nn1355, https://doi.org/10.1038/nn1355.

Klein, I., Dubois, J., Mangin, J.-F., Kherif, F., Flandin, G., Poline, J.-B., ... Bihan, D. (2004). Retinotopic organization of visual mental images as revealed by functional magnetic resonance imaging. *Cognitive Brain Research*, *22*(1), 26–31, https://doi.org/10.1016/j.cogbrainres.2004.07.006.

Kosslyn, S., Alpert, N., Thompson, W., Maljkovic, V., Weise, S., Chabris, C., ... Buonanno, F. (2007). Visual mental imagery activates topographically

organized visual cortex: PET investigations. *Journal of Cognitive Neuroscience*, *5*(3), 263–287, https://doi.org/10.1162/jocn.1993.5.3.263.

Kosslyn, S., Ganis, G., & Thompson, W. (2001). Neural foundations of imagery. *Nature Reviews Neuroscience*, *2*(9), 635–642, https://doi.org/10.1038/35090055=.

Kosslyn, S. M., Pascual-Leone, A., Felician, O., Camposano, S., Keenan, J. P., Thompson, W. L., ... Alpert, N. M. (1999, April 2). The role of area 17 in visual imagery: Convergent evidence from PET and rTMS. *Science*, *284*(5411), 167–170, https://doi.org/10.1126/science.284.5411.167.

Kosslyn, S., Thompson, W., & Alpert, N. (1997). Neural systems shared by visual imagery and visual perception: A positron emission tomography study. *NeuroImage*, *6*(4), 320–334, https://doi.org/10.1006/nimg.1997.0295.

Kosslyn, S. M., Thompson, W. L., Kim, I. J., & Alpert, N. M. (1995, November 30). Topographical representations of mental images in primary visual cortex. *Nature*, *378*(6556), 496–498, https://doi.org/10.1038/378496a0.

Kowler, E. (1989). Cognitive expectations, not habits, control anticipatory smooth oculomotor pursuit. *Vision Research*, *29*(9), 1049–1057, https://doi.org/10.1016/0042-6989(89)90052-7.

Kowler, E., Aitkin, C., Ross, N., Santos, E., & Zhao, M. (2014). Davida Teller Award Lecture 2013: The importance of prediction and anticipation in the control of smooth pursuit eye movements. *Journal of Vision*, *14*(5):10, 1–16, https://doi.org/10.1167/14.5.10. [PubMed] [Article]

Kowler, E., & Steinman, R. (1979a). The effect of expectations on slow oculomotor control—I. Periodic target steps. *Vision Research*, *19*(6), 619–632, https://doi.org/10.1016/0042-6989(79)90238-4.

Kowler, E., & Steinman, R. (1979b). The effect of expectations on slow oculomotor control—II. Single target displacements. *Vision Research*, *19*(6), 633–646, https://doi.org/10.1016/0042-6989(79)90239-6.

Kowler, E., & Steinman, R. (1981). The effect of expectations on slow oculomotor control—III. Guessing unpredictable target displacements. *Vision Research*, *21*(2), 191–203, https://doi.org/10.1016/0042-6989(81)90113-9.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (2008). Backpropagation applied to handwritten zip code recognition. *Neural Computation*, *1*(4), 541–551, https://doi.org/10.1162/neco.1989.1.4.541.

Nguyen, A., Yosinski, J., & Clune, J. (2014). Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. *arXiv*1412.1897.

Noton, D., & Stark, L. (1971). Scanpaths in saccadic eye movements while viewing and recognizing patterns. *Vision Research*, *11*(9), 929–I N8, https://doi.org/10.1016/0042-6989(71)90213-6.

Pasternak, T., & Greenlee, M. (2005). Working memory in primate sensory systems. *Nature Reviews Neuroscience*, *6*(2), nrn1603, https://doi.org/10.1038/nrn1603.

Rawat, W., & Wang, Z. (2017). Deep convolutional neural networks for image classification: A comprehensive review. *Neural Computation*, *29*(9), 1–98, https://doi.org/10.1162/neco_a_00990.

Springer, A., de Hamilton, A., & Cross, E. (2012). Simulating and predicting others' actions. *Psychological Research*, *76*(4), 383–387, https://doi.org/10.1007/s00426-012-0443-y.

Springer, A., Parkinson, J., & Prinz, W. (2013). Action simulation: Time course and representational mechanisms. *Frontiers in Psychology*, *4*, 387, https://doi.org/10.3389/fpsyg.2013.00387.

Stark, L., Vossius, G., & Young, L. R. (1962). Predictive control of eye tracking movements. *IRE Transactions on Human Factors in Electronics*, *HFE-3*(2), 52–57, https://doi.org/10.1109/THFE2.1962.4503342.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2013). Intriguing properties of neural networks. *arXiv*1312.6199.

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, https://doi.org/10.1016/j.tics.2017.05.012.

Umiltà, M. A., Kohler, E., Gallese, V., Fogassi, L., Fadiga, L., Keysers, C., & Rizzolatti, G. (2001). I know what you are doing. A neurophysiological study. *Neuron*, *31*(1), 155–165, https://doi.org/10.1016/S0896-6273(01)00337-3.

Winawer, J., Huk, A., & Boroditsky, L. (2010). A motion aftereffect from visual imagery of motion. *Cognition*, *114*(2), 276–284, https://doi.org/10.1016/j.cognition.2009.09.010.

Yamins, D., Hong, H., Cadieu, C., Solomon, E., Seibert, D., & DiCarlo, J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, *USA*, *111*(23),

8619–8624, https://doi.org/10.1073/pnas.1403112111.

Yosinski, J., Clune, J., Nguyen, A., Fuchs, T., & Lipson, H. (2015). Understanding neural networks through deep visualization. *arXiv*1506.06579.

Supplementary material

Supplementary Movie S1.
Supplementary Movie S2.