Physical predictions over time

Kevin A Smith (k2smith@ucsd.edu),[1] Eyal Dechter (edechter@mit.edu),[2]
Joshua B Tenenbaum (jbt@mit.edu),[2] Edward Vul (evul@ucsd.edu)[1]
1. University of California, San Diego, Department of Psychology, La Jolla, CA 92093
2. MIT, Department of Brain and Cognitive Sciences, Cambridge, MA 02139

Abstract

In order to interact with the world, people must be able to predict how it will unfold in the future, and these predictions must be updated regularly in light of new information. Here we study how the mind updates these predictions over time. Participants were asked to make ongoing predictions about the destination of a simulated ball moving on a 2D bumper table. We modeled these decisions by assuming people simulate the world forward under uncertainty. This model fit participants' behavior well overall, suggesting that people continuously update their physical simulations to inform their decisions. In some specific scenarios participants' behavior is not fit well by the simulation based model in a manner suggesting that in certain cases people may be using qualitative, rather than simulation-based, physical reasoning.

Keywords: intuitive physics; forward simulation

Introduction

Changing lanes while driving seems like a simple and ordinary task – millions of people do it everyday. But to do so safely requires sophisticated predictions. Drivers must judge where their own car and those around it will be during the lane change, and, crucially, they must update these predictions with new information: if a car in the adjacent lane accelerates, a driver may abort her lane change to avoid a collision.

This scenario demonstrates how people typically plan their actions: prediction is updated as new information is gathered. Research spanning decades has investigated how people predict future object movement while objects are hidden (Faisal & Wolpert, 2009; Rosenbaum, 1975; Runeson, 1975; Smith & Vul, 2013; Téglás et al., 2011), but in most natural cases, observers continue to see objects while updating their predictions. In this study we investigate how people change their instantaneous predictions about objects over time: are ongoing predictions the result of online simulation?

Recent research provides evidence that people use 'Noisy Newtonian' models of physics to simulate the world (Sanborn, Mansinghka, & Griffiths, 2013): peoples' internal physical models are based on correct assumptions about physics, but uncertainties in object position, movement, and latent variables can cause biases and variability in prediction. This framework has been used to predict peoples' judgments about the stability of a tower of blocks (Hamrick, Battaglia, & Tenenbaum, 2011), the movement of hidden objects (Smith & Vul, 2013), and even judgments about physical causality (Gerstenberg, Goodman, Lagnado, & Tenenbaum, 2012). These works, however, solicited predictions at single instances in time. In this paper, we investigate whether a model that assumes faithful physics under uncertainty is also consistent with how peoples' predictions evolve over time. We show that people's decisions are often consistent with online forward simulation, but we also find that people can use qualitative reasoning about the world (e.g., Forbus, 1994) when this is more informative than simulations.

Experiment

We asked participants to play a game in which they make predictions about the path of a ball bouncing around a computerized table. The ball can reach one of two targets on the table, and participants earn points for predicting which target it reaches first. Crucially, they make this prediction continuously throughout the trial, earning points while predicting the correct target but losing points while predicting the incorrect target. In this way, we could capture how uncertainty (decisions whether to choose a target) and choices (which target) evolved over the course of each trial.

Methods

Sixty-six UC San Diego undergraduates participated in this experiment for course credit.[1]

On each trial, participants saw a ball moving around a 'table' on the computer screen that contained blocks and both a red and a green target. The ball bounced perfectly elastically off of the edge of the table and blocks, ending when the ball reached one of the two targets. While the trial progressed, participants were asked to predict whether the ball would hit the red target or the green target first, indicating their guess by holding down either the 'z' or the 'm' key (each key counterbalanced for red and green between participants). If they were unsure, participants could press neither key, and if their prediction changed mid-trial, they were encouraged to switch keys. Holding down a key would fill a bar of the associated color, and at the end of the trial, the score would be determined by the difference between the proportion of time the keys for each target were held down:

$$Score = 20 + 100 * (Prop(Correct) - Prop(Incorrect)) \quad (1)$$

After each trial, participants were notified of their score and could continue to the next trial by pressing the spacebar.

Participants were each given the same 400 trials in a random order. Of these, 370 trials were randomly generated, and 30 were designed to consider various extreme scenarios.

[1] We excluded one participant for holding down a single key through the entirety of the second half of the experiment.

Of special note in the hand-crafted trials are five trials in which the configuration of the walls made it impossible for the ball to ever reach one of the two targets; these are called the ‘qualitative’ trials as they were meant to differentiate between simulation-based intuitive physics and a qualitative assessment of the table configuration.

Each trial lasted between 2.0s and 10.2s. Target colors were randomly swapped for each trial to avoid color bias effects.[2] Responses were polled and recorded once every tenth of a second.

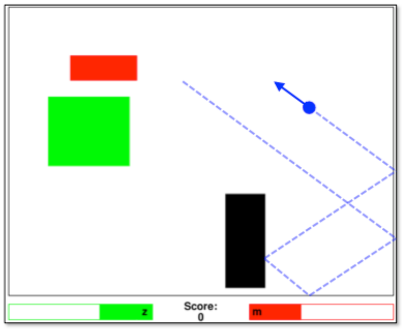

Figure 1: Illustration of an ongoing trial. Participants would see the ball travel along the dotted line and would predict green or red during its motion (neither blue line was visible).

Results

We analyzed participants’ aggregate performance across trials via their total score (eq. 1). Participants showed low variability in their average trial scores (mean = 56.0, sd = 5.1) and scores for each trial were very consistent across participants (split half correlation, r = 0.96).

We investigated whether there were surface-level features of trials that make them more or less difficult. Here we use average score as a proxy for difficulty; high scores indicated that most participants could accurately predict the path of the ball easily (and early), while low scores indicated uncertainty and mis-prediction.

The features we considered as possible predictors included: (1) trial duration, (2) the number of blocks on the table, (3) the number bounces before the ball hit the target, (4) the initial deviation of the ball’s path from a horizontal or vertical direction, (5) the proportion of the table clear of walls or targets, (6) the ratio of the area of the correct target to the incorrect target, (7) the ratio of the average distance of the ball to the correct target versus the incorrect target, and (8) the closest the ball ever was to the incorrect target.

We found four predictive features: trials were easier when their trajectory was on average closer to the final target, involved fewer bounces, when the initial motion was along a cardinal direction, and when the ball never approached the incorrect target. These four predictors together explained 31.9% of the variance in scores across trials.

Table 1: Predictors of trial difficulty. Partial correlation is correlation between predictor and trial score accounting for all other predictors.

Predictor	r	$r_{partial}$
(7) Distance ratio	-0.51	-0.24
(3) Bounces	-0.25	-0.19
(4) Direction deviation	-0.21	-0.14
(8) Nearness to incorrect	0.42	0.09

Although, aggregate metrics of ball trajectory accounted for some of the variation in difficulty across trial, such an analysis fails to capture the rich predictions individuals make over time and how those change in light of the details of a given table configuration and ball trajectory.

To further delve into how people make online physical predictions and explain why some of these features might make trials more difficult, we compared human behavior to predictions made via stochastic physical reasoning.

Physical Prediction Model

Description

The model we used to predict behavior on this task has two parts: the physical simulator, which provides possible paths that the ball can take, and the decision policy, which uses the output of the physical simulations to decide which target to choose (if either).

Physical simulator The part of the model that simulates the trajectory of the ball is based in large part on the model of Smith & Vul (2013). This model assumes that people base their physical models on real physics but must incorporate uncertainty about the world into their physical judgments.

This model captures two sources of uncertainty: 1) *perceptual uncertainty* arises from the noisiness of inferring the position and movement of objects, and 2) *dynamic uncertainty* is uncertainty about the roughness and elastic properties of the table and walls that could cause the ball’s path to deviate from idealized Newtonian physics over time.

The physical simulator produces 500 simulation paths[3] every tenth of a second for each trial to replicate the polling frequency in the experiment. These simulation paths were produced using the same uncertainty parameters and fits as Smith & Vul (2013).[4] Each path terminates when the simulated ball reaches either the red or the green target, or when 10 seconds of simulated time has passed.[5]

[2] Individual responses were corrected in swapped trials to allow for consistent analysis.

[3] We used 500 simulations to estimate the aggregate distribution of predictions over all individuals. We suspect that each individual used far fewer simulations, but aggregate across-subject behavior is consistent with many simulated paths over the full set of subjects (see Teglas et al. for a related discussion).

[4] Due to computational limitations, these parameters could not be easily fit to this data. However, because they were fit to aggregate behavior in the prior model, we assumed that they would capture aggregate performance in this similar task as well.

[5] This was an upper bound nearly equivalent to the longest trial.

The physical simulator outputs a set of proportions: how many paths reached the green target, how many reached the red target, and how many did not reach a target within the maximum simulation time (‘uncertain’ paths).

Decision policy The decision policy takes the output of the physical simulator and assigns belief to two decisions: 1) is there sufficient certainty about which target the ball will hit to offer any guess at all? (analogous to participants’ decision whether or not to press any button at all), and if so, 2) which target should be guessed?

The decision of whether any prediction should be made is based on the proportion of simulation paths that reached either target. These are combined and fit with two parameters: a parameter α representing a Luce choice soft-max weighting, and a parameter γ to capture a bias towards making any guess at all:

$$(2) \qquad P(Any) = \frac{Sim(Red\ or\ Green)^{\alpha} + \gamma}{Sim(Red\ or\ Green)^{\alpha} + Sim(Uncertain)^{\alpha} + \gamma}$$

Conditioned on the decision to make any guess at all, the decision whether to guess red or green is based on the relative proportion of simulated paths that reached each color target. Here, there is a single Luce choice soft-max weighting parameter (β), but because experimental trial colors were randomized, we assumed no bias:

$$(3) \qquad P(Red|Any) = \frac{Sim(Red)^{\beta}}{Sim(Red)^{\beta} + Sim(Green)^{\beta}}$$

Finally, we assume there is a decision offset: people cannot immediately use simulation information, but instead must take time to process it, come to a decision, and move their hand to push a button. Thus we fit a single parameter *t* to determine how long the model should wait to use simulation information.[6]

These four parameters were optimized to fit the empirical data at each polling time point for each trial. At each point, we created a vector of the empirical probabilities of pressing each of the two keys at time point i of trial j: [$prop(Red)_{ij}$, $prop(Green)_{ij}$]. We then calculated a similar vector of model predictions: [$P(Red)_{ij}$, $P(Green)_{ij}$]. The parameters were fit to minimize the total Euclidean distance between these points over all time points of all trials:

$$(4) \quad \sum_j \sum_i \sqrt{(prop(Red)_{ij} - p(Red)_{ij})^2 + (prop(Green)_{ij} - p(Green)_{ij})^2}$$

Model performance

How do predictions change over time? Participants often changed their decisions as trials progressed, either from uncertain to certain or one color to the other. We first ask whether we can capture changes in participants’ predictions over time. We compared model decision probabilities to the distribution of participants’ choices at each time point: what proportion of participants guessed the ball would end in the red target or the green target, or were too uncertain to offer a guess. Although there were differences in individual predictions, we believe aggregation is appropriate given the low variability in total scores and high consistency within trials (split half correlation, r = 0.96).

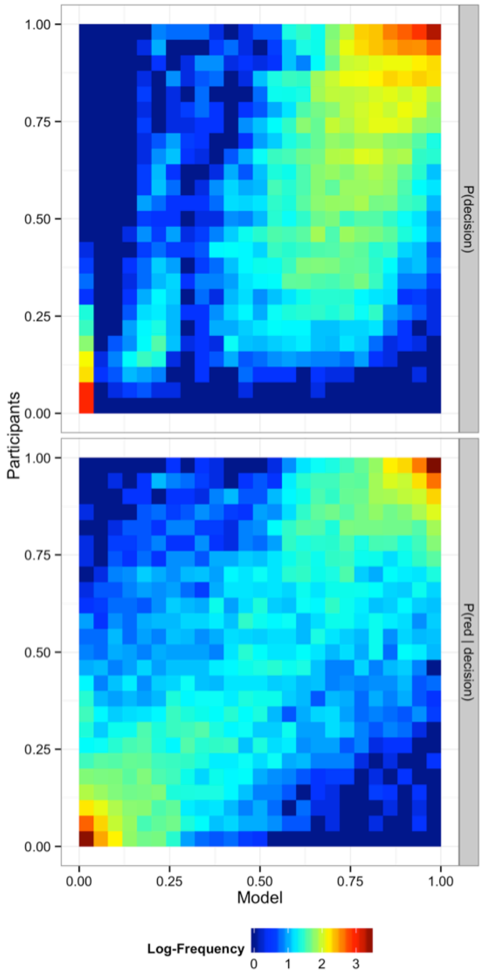

Figure 2: Joint histogram of model (x-axis) and human (y-axis) decisions. (Top) The probability of making any guess (pushing a button). (Bottom) The probability of choosing ‘red’ given a decision. Colors indicate log-frequency of time points in each bucket, with hotter colors indicating more observations. Observations along the diagonal indicate the model is accurately capturing the exact proportions of participants making decisions.

If people make decisions based on similar uncertainty and decision policies to those of the model, then the model should be able to predict both (a) when people make any decision and (b) which choice they make (red or green) when they do. Figure 2 shows the correlation between model predictions and participants’ behavior. In the top panel, we see that participants’ decisions whether to push either button are well predicted by the model (r = 0.84): at most time points either our model believes that no guess should be made and nearly no participants offer a guess

[6] Although the model only simulated once every tenth of a second, this parameter could take on continuous values. If it fell between two simulation times, the model decision would be a weighted average of each of those two decisions.

(bottom-left), or our model believes that a guess should definitely be made and nearly all participants offer a guess (top-right). Moreover, even when participants are not unanimous in their decision to offer a guess, the model captures the variation in the proportion of people making guesses. In the bottom panel, we see that the choice people make (red or green) at time points when they do offer a guess, is well explained by the model (r = 0.92): again much of the time participants are nearly unanimous in their choice of one of the targets, as is the model. But again the model also captures the gradations in beliefs when participants are split on which target to choose

What makes trials difficult? We also wanted to know if we could better explain what makes trials easier or harder. To do so, we calculated the average model score for each trial in an equivalent way to participants' scores:

$$(5)\ ModelScore = 20 + 100 * \sum_t (P(correct)_t - P(incorrect)_t)$$

The model predicts the difficulty of the trials better ($R^2 = 0.675$; see Figure 3) than using superficial trial features, as we first investigated ($R^2 = 0.319$). It is noteworthy that our model was never explicitly informed about how the trial would unfold, which characteristics should make a trial more difficult, or even how scoring works; nor was the model fit to capture trial scores. Instead simply by considering variations in moment-by-moment physical predictions, we could capture variation in trial difficulty.

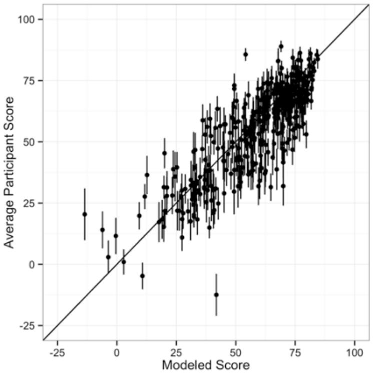

Figure 3: Modeled versus empirical trial scores. Each point represents a single trial, where bars are 95% confidence intervals on empirical scores.

There remains reliable variability in participants' average scores that is not explained by the model. To investigate what might be causing this, we again predicted participants' scores on each trial, using the same features that we had before, but physical model's score included as a predictor. With the model score added, no new features became significant predictors (indicating that the model is unbiased with respect to those features), and two features – the number of bounces and the smallest distance to the incorrect target – were no longer good predictors (indicating that the model accounts for these difficulties well). Two features did remain though: the deviation of the path from the horizontal or vertical, and the ratio of the average distance of the ball to each of the targets. Including these two predictors did provide a statistically significantly better fit (*F(2,393)=12.9, p<0.001*), but only explained slightly more variability in participants' scores ($R^2 = 0.695$).

The remaining feature predictors inform us about what aspects of human cognition the model is not capturing. First, the model does not capture the additional difficulty introduced when the ball is traveling at an angle. The physical model assumes that directions of movement are equally difficult to simulate, but this indicates that people may find it difficult to predict the path of objects travelling at an angle, in the same way people have difficulty discriminating oblique motion (Matthews & Quin, 1999). The ratio of the distances between the targets also remains as a predictor, but the correlation is attenuated as compared to the relationship without the model predictions, perhaps suggesting that there is a bias to believe the ball will hit the nearer target beyond simulations.

Table 2: Predictors of trial difficult including the physical model. Partial correlation is correlation between predictor and trial score after all other predictors have been included. Prior partial correlation is partial correlation of prediction not including the physical model (see Table 1)

Predictor	r	$r_{partial}$	Prior $r_{partial}$
Physical model	0.82	0.65	N/A
(4) Direction deviation	-0.21	-0.17	-0.14
(7) Distance ratio	-0.51	-0.15	-0.24

Why do human and model predictions differ? Participants' predictions were overall consistent with the model, but this fit varied by trial. To explain how people might be consistent with or deviate from simulation using noisy physics, we investigated how well the model fit empirical data on each trial. Our metric of trial fit was the average deviation between participants' decisions and model predictions over the trial, similar to eq. 4:

$$(6)\ Dev_j = \sum_i \sqrt{(prop(Red)_{ij} - p(Red)_{ij})^2 + (prop(Green)_{ij} - p(Green)_{ij})^2} / t_j$$

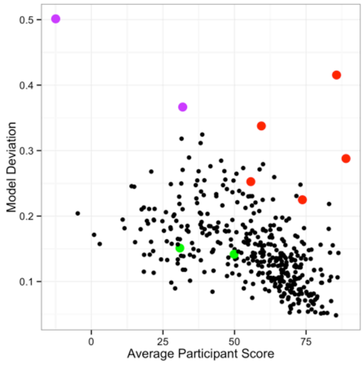

Figure 4: Model deviation from empirical decisions as a function of trial difficulty.

Well-fit trials had lower average deviations, whereas poor fits were characterized by high deviation. As can be seen in

Figure 4, the model predicted participants' decisions better on trials that the participants found easy, with a slight reduction in performance as the trials became more difficult.

Many of the highest scoring and best fitting trials were straightforward, as people quickly decided on the correct target, and the model captured this behavior. We first review two trials with average model fit (the green points, Figure 4) to discuss the strengths of the model, then review where the model deviates from human performance: two trials that people find difficult but the model does not (the purple points), and the five 'qualitative' trials that the model finds difficult but people do not (the red points). Information on all other trials can be found online at experiments.evullab.org/physovertime/trials.html

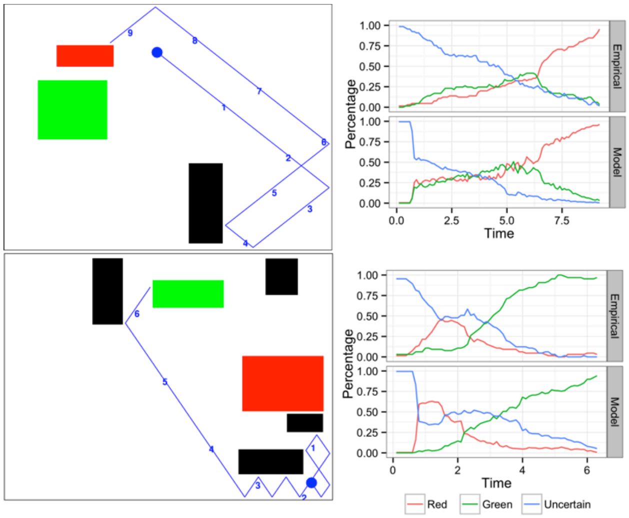

Figure 5: (Left) Image of average fitting trials. Numbers represent the time in seconds when the ball passes that point. (Right) Associated proportion of red, green, or undecided decisions by participants (top) or the model (bottom). Time on the x-axis is matched to the associated point in the trial diagram. The model tends to capture human behavior, erring only in confidence (top trial) or timing (bottom trial).

Moderately difficult trials were those in which the correct target was not immediately obvious, requiring participants to either resolve their belief over time (see Figure 5-top), or change their beliefs when more information arrived (see Figure 5-bottom). The model predicted these types of decisions well, typically erring only in either the amount of uncertainty or the timing of decision changes. This suggests that in aggregate, the model captures the way that people resolve uncertainty: certainty increases as the ball travels or when people see the outcome of a bounce.

There were also trials that participants did poorly on that the model did not fit well. These were typically trials where (a) the ball was traveling at a steep angle, and (b) small changes in the perceived layout of the table could cause a difference in the ending target. For instance, the top trial in Figure 6 was the trial for which the model performed the worst. Here, if the wall just below the green target were slightly larger, the ball would miss that target and hit red. The model assumes perfect knowledge of the table, but it is likely that people have uncertainty about the area of the bumpers and target-areas as well – uncertainty that the model does not have.

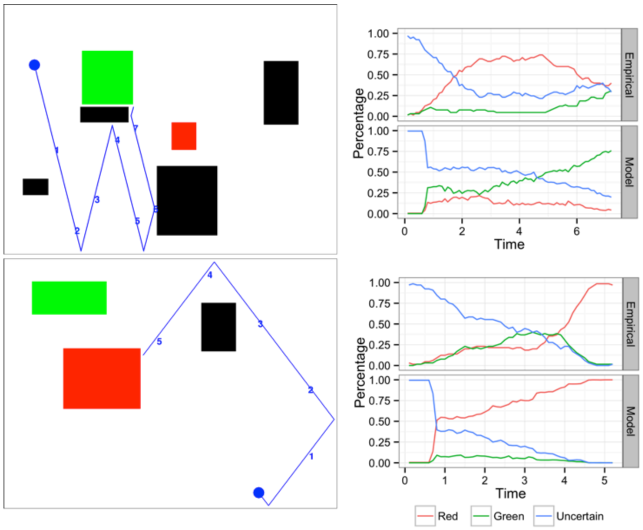

Figure 6: Image of trial path (left) and associated empirical and model predictions (right) for difficult, poor-fitting trials. Slight changes in the layout of the table would have significant consequences for both trials, which the model does not capture.

We also specifically created 'qualitative' trials in which one target was difficult for the ball to get to, but the other was unreachable (see Figure 7). These trials comprised a large portion of the trials that the model found difficult but participants found easy. If participants are deciding between the two targets based solely on the output of a physical simulator, then they should show large amounts of uncertainty until near the end of the trial. On the other hand, if people can qualitatively analyze the structure of the table and determine that one target is unreachable, then they should quickly show high confidence for the possible but unlikely target. On these trials, participants tended to be much more confident than model predictions starting early in the trial, suggesting that they were performing a qualitative, topological analysis of the possible trajectories and outcomes on the table.

Discussion

In this study we demonstrated that human online predictions about the world can be well captured by a model that continually simulates the world forward using noisy physical principles. This physical simulation model could better predict which trials were easy and which were difficult than a simple analysis of trial features. Likewise, it well predicted how often people would decide on one of the two target, and which target they would decide on. Together, these results suggest that people are performing forward physical simulation online as the trial unfolds.

Furthermore, we found aspects of human behavior that the model missed, for instance difficulty with balls traveling at oblique angles or uncertainty about the state of the table. These factors are consistent with a forward simulation model, and suggest refinements to our knowledge of how people simulate object movement.

The qualitative trials, on the other hand, are inconsistent with a pure simulation account. On these trials, simulations would have difficulty reaching the incorrect target, and thus pure simulation would predict uncertainty throughout the trial, as our model does. Instead, people quickly grow certain, reasoning that if the ball cannot reach one target, it must eventually reach the other. This suggests that people may use qualitative spatial reasoning (e.g. Forbus, 1994) when it provides a clear answer about their environment, but otherwise use simulations to reason about the future. Further investigation is required to understand how and when people decide to switch between modes of prediction.

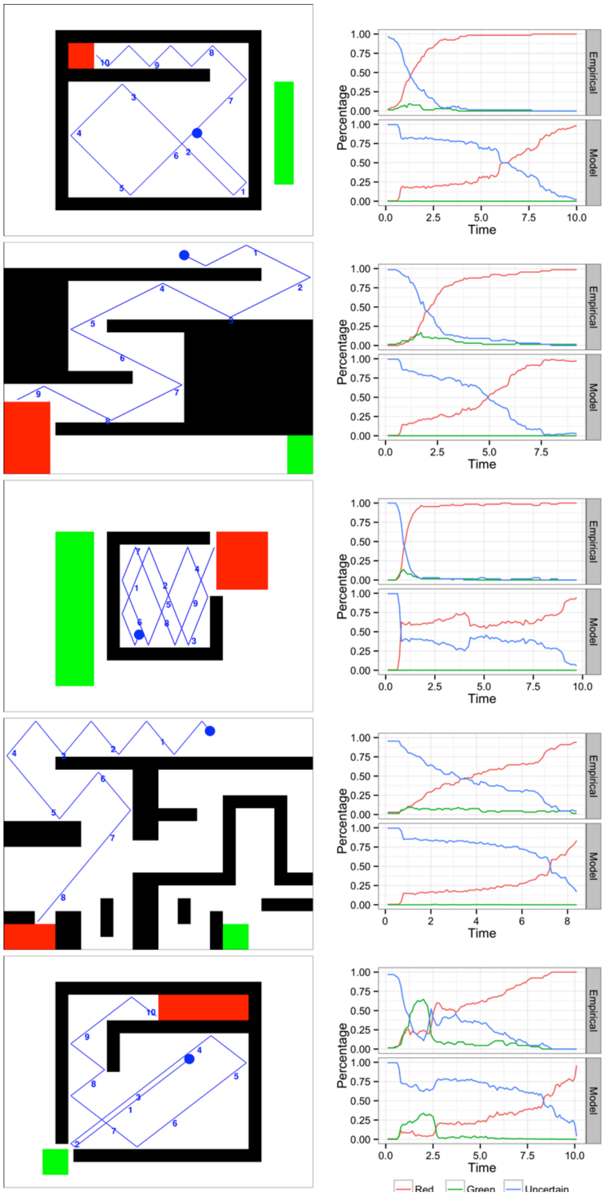

Figure 7: Image of trial path (left) and associated empirical and model predictions (right) for qualitative trials. Participants can quickly discover that the ball cannot reach one of the two targets and select the other, but the model must use simulations to make this determination.

Finally, our model made predictions by generating a new set of simulated paths at each time step. However, people probably conserve computation and only slightly update a single set of simulations at each time step (e.g., using a particle filtering algorithm). Details of the exact prediction algorithm are a fruitful area for future research.

We can predict the future state of the world and integrate new information to make better predictions. This study suggests these predictions often come from continuously updating our vision of how the world will unfold, but also leaves tantalizing clues that we can overlay our simulations with qualitative reasoning about our environment as well.

Acknowledgments

KS and EV were supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of the Interior (DOI) contract D10PC20023. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DOI, or the U.S. Government.

References

Faisal, A. A., & Wolpert, D. M. (2009). Near optimal combination of sensory and motor uncertainty in time during a naturalistic perception-action task. *Journal of Neurophysiology, 101*, 1901-1912.

Forbus, K. D. (1994). *Qualitative spatial reaonsing: Framework and frontiers*.

Gerstenberg, T., Goodman, N., Lagnado, D. A., & Tenenbaum, J. B. (2012). *Noisy Newtonians: Unifying process and dependency accounts of causal attribution*. Paper presented at the Proceedings of the 34th Annual Meeting of the Cognitive Science Society, Sapporo, Japan.

Hamrick, J., Battaglia, P., & Tenenbaum, J. (2011). *Internal physics models guide probabilistic judgments about object dynamics*. Paper presented at the Proceedings of the 33rd Annual Meeting of the Cognitive Science Society, Boston, MA.

Matthews, N., & Quin, N. (1999). Axis-of-motion affects direction discrimination, not speed discrimination. *Vision Research, 39*, 2205-2211.

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

Runeson, S. (1975). Visual prediction of collision with natural and nonnatural motion functions. *Perception and Psychophysics, 18*(4), 261-266.

Sanborn, A. N., Mansinghka, V. K., & Griffiths, T. L. (2013). Reconciling intuitive physics and Newtonian mechanics for colliding objects. *Psychological Review, 120*(2), 411-437.

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

Téglás, E., Vul, E., Girotto, V., Gonzalez, M., Tenenbaum, J. B., & Bonatti, L. L. (2011). Pure reasoning in 12-month-old infants as probabilistic inference. *Science, 332*, 1054-1059.