www.sciencemag.org/cgi/content/full/332/6033/1054/DC1

Supporting Online Material for

Pure Reasoning in 12-Month-Old Infants as Probabilistic Inference

Ernő Téglás, Edward Vul, Vittorio Girotto, Michel Gonzalez, Joshua B. Tenenbaum,*
Luca L. Bonatti*

*To whom correspondence should be addressed. E-mail: lucabonatti@mac.com (L.B.); jbt@mit.edu (J.B.T.)

Published 27 May 2011, *Science* **332**, 1054 (2011)
DOI: 10.1126/science.1196404

This PDF file includes:

Materials and Methods
Figs. S1 to S8
References

Other Supporting Online Material for this manuscript includes the following:
available at www.sciencemag.org/cgi/content/full/332/6033/1054/DC1

Movies S1 to S5

Pure reasoning in 12-month-old infants as probabilistic inference: Supporting Online Material

Ernő Téglás [1,2#] Edward Vul [3#], Vittorio Girotto [4,5], Michel Gonzalez [5], Joshua B. Tenenbaum [6] and Luca L. Bonatti [7]

1 Cognitive Development Centre, CEU, H-1015, Budapest, Hungary
2 Hungarian Academy of Sciences, Budapest, Hungary
3 University of California, San Diego
4 University IUAV of Venice, Italy.
5 CNRS and and Aix-Marseille University, France.
6 Massachusetts Institute of Technology, Cambridge, MA
7 ICREA and Universitat Pompeu Fabra, Barcelona, Spain

# These authors contributed equally to this work.

Address correspondence to:

Luca L. Bonatti
ICREA and Universitat Pompeu Fabra
C. Roc Boronat, 138
Edifici Tanger, 55.110
08018 Barcelona, ES
and
Joshua B. Tenenbaum
Department of Brain & Cognitive Science
Massachusetts Institute of Technology
77 Massachusetts Ave,
Cambridge, MA 02139, USA

email: lucabonatti@mac.com, jbt@mit.edu

Supporting Online Movies

Movies S1, S2, S3, S4, S5

Movie S5: Example of a familiarization movie

Movie S1: low probability exit, low probability close, 2s occlusion

Movie S2: low probability exit, high probability close, 1s occlusion

Movie S3: high probability exit, high probability close, 0s occlusion

Movie S4: high probability exit, low probability close, 0s occlusion

Experiments

Experiment 1

Participants

Twenty 12-month-old infants (M= 12.18m, range 12m 6d- 13m 1d) participated in the experiment (8 boys). An additional 12 infants were excluded from analysis (4 were fussy, 3 cried, 3 exceeded maximum looking time criteria, looking more than 30 s in two or more test trials, 1 looked away exactly when the ball exited the frame, and 1 was sleepy). Parents of the infants gave informed consent before the experiment.

Stimuli and Material

We generated QuickTime movies as 3D animations with Maya 8.5, at 25 fps. They represented two kinds of solid objects with different shapes (star-like/cube-like) and colors (yellow/blue) bouncing inside a container, as in a lottery game, with an open exit in its lower part. The objects' motion was determined by identical physical parameters.

There were two different types of movies, one for the familiarization and the other one for the test phase. The familiarization movies presented two objects of each kind bouncing for 14.5 s. After the bouncing period, one object exited the container, thus showing infants the outcome of interest. After the exit, an occluder progressively covered the container and its content, and the trial ended. The length of the familiarization movies was 19.24 s.

The test movies were constructed like the familiarization movies, with two crucial differences. First, three identical objects and one different object bounced inside the container. Second, the temporal order of the exit phase and the occlusion phase were inverted. Thus, at the end of the bouncing period (13 s), the occluder progressively faded over the scene (2 s); then a full occlusion period lasting 2 s followed, and finally an object exited the container. After the exit, the occlusion disappeared, leaving all objects visible again. The configuration visible before the occlusion was controlled in such a way that the object(s) of one kind was (were) always far from the exit, while the object(s) of the other kind was (were) close to it. Across the test movies, the distance between the object sets reached its maximum at 0.9 s (SD=0.12 s) before the full occlusion. Sixteen test movies were constructed from all possible combinations of class numerosity of the outcome (3-instances/1-instance), distance to the exit before occlusion (near/far), color (blue/yellow) and shape (star-like/cube-like).

The experiment was run with the software PsyScope X *(S1)* (http://psy.cns.sissa.it), on an Apple DualG5 computer. Movies were presented on a 17 in screen, with the object frame covering a 14x14 cm area. A camera hidden behind the screen digitally recorded infants' faces on a separate Apple computer, as an iMovie project. Movies were then inspected offline with the software PsyCode (http://psy.cns.sissa.it/) to extract exact looking time durations.

Procedure

Infants sat on their caretaker's laps, approximately 80 cm from the screen, in a darkened room. Caretakers wore black opaque glasses during the experiment. They were instructed not to interact with infants. They were invited to hold infants at their hips with both hands and let them move freely, but to not let infants abandon their initial posture (facing the screen) for more than 5 s, after which period they had to gently turn them back towards the center. The experimenter, who was blind to the experimental conditions, monitored infants' behavior from a separate screen. Before each movie, a visual attractor helped orienting infants' attention towards the centre. When infants looked at the centre, a movie started playing. To ensure that every infant saw every movie in its entirety, the presentation of the stimuli was infant-controlled. Movies were paused when infants were not paying attention and continued playing when they looked back at the screen.

Infants were presented with two familiarization trials, in which the movie ended with an occlusion and no looking time was measured. Immediately after familiarization, four test trials began, each one presenting a different experimental movie. Lists with different orders of experimental movies were presented, half of which had a probable and half an improbable trial first. In the lists, shapes and colors were counter-balanced using a latin-square design. Every participant was presented all the four possible combinations of probability and distance. The order of presentation was counterbalanced across participants.

In the test trials, at the end of the occlusion phase, a sound marked the outcome. When infants looked at the center, the exiting phase (with an object exiting the container) started, and looking time monitoring began. A trial ended if infants looked away for more than 2 consecutive s, or looked for more than 30 cumulative s. Looking time was recoded off-line for further analysis. Infants were excluded from the analysis if they had cumulative timeouts in two or more trials, or if they fussed out. Trials were also excluded when infants turned away exactly when the object exited the container.

<u>Results and Discussion</u>

Data exceeding 2 SD were filtered out (2 out of 80 data points). Repeated measures ANOVAs with class numerosity and distance as within-subject factors showed that infants looked longer at the outcomes in which the single object exited the container ($F(1, 19) = 5.66$, $P = 0.028$) but revealed no other main effect or interactions. This pattern was found in 14 out of the 20 participants ($P = 0.1$, binomial test, two-tailed). We tested with separate ANOVAs if the perceptual properties of the exiting object contributed to the result. No effect of color (M_{Yellow}= 14.1 s, M_{Blue}=13.4 s; $F(1,19) =0.3$, $P = 0.5$) or shape (M_{Cube}=13.2 s, M_{Star}= 14.2 s; $F(1,19) = 0.5$, $P = 0.4$) was found.

A mixed ANOVA with class numerosity (1-instance/3-instances), distance (near/far) as within-participant factors and order of presentation (3-instances first /1-instance first) as between-participant factor revealed a trend towards and effect of order or presentation. Infants tended to look longer when the test trials started with an improbable outcome ($M_{1\text{-instance First}}$ = 15.6 s; $M_{3\text{-instances First}}$ = 11.9 s; $F(1,52) = 4.09$, $P = 0.058$). Order of presentation interacted with class numerosity ($F(1,52) = 4.76$, $P = 0.03$). Post-hoc analysis showed that when infants saw an outcome with the single different object out first, they looked longer at 3-class outcomes ($M_{1\text{-instance}}$ = 19.1 s, $M_{3\text{-instances}}$ = 12 s; Scheffe post-hoc, $P = 0.001$), but not so when they saw a 3-class outcome first ($M_{1\text{-instance}}$= 12 s, $M_{3\text{-instances}}$ = 11.7 s, Scheffe post-hoc, $P = 0.8$). In similar experiments *(S2*, e.g., Experiment 1*)*, and in several other experiments run in our laboratory, no such order effect was found, suggesting that the current trend is due to group variability. Nevertheless, although an order effect was not predicted, such an effect may be interpretable. In adult literature using probability matching paradigms, adult participants have produced shorter reaction times after a miss than after a hit *(S3)*. The phenomenon has been interpreted as a consequence of an error monitoring process. Participants may become more likely to be aware to the properties of the stimulus after a miss than after a hit. It is not impossible that a similar phenomenon occurs in infants, who may be more attentive to a stimulus after an unexpected outcome (the equivalent of a miss) than after an expected outcome (the equivalent of a hit). Confirmation of this interpretation would require further research.

Experiment 2

Participants

Twenty 12-month-old infants (M= 12.22m, range 12m 6d - 13m 5d) participated in the experiment (12 boys). An additional 19 infants were excluded from analysis (11 fussed out, 4 had two or more cumulative timeouts, 2 looked away exactly when the ball exited, 1 for technical problems, 1 for caregivers' intervention during the experiment).

Stimuli, material and Procedure

Material and procedure were identical to Experiment 1, except that the full occlusion period in the experimental movies lasted 1 s.

Results and Discussion

Data exceeding 2 SD were filtered out (4 out of 80 data points). Repeated measure ANOVA with class numerosity and distance as within subject factors confirmed that infants looked longer at the 1-class object outcomes ($F(1, 19) = 4.65$, $P = 0.04$), but they also looked longer when the object which was farther from the exit before occlusion exited the container ($F(1, 19) = 5.22$, $P = 0.03$). Distance and class numerosity did not interact ($F(1, 17) = 2.09$, $P = 0.17$). Neither the color (M_{Blue}= 13.2 s, M_{Yellow}= 13.1 s; $F(1,19)=0.01$, $P = 0.9$) nor the shape (M_{Cube}= 12.8 s, M_{Star}= 13.5 s; $F(1, 19) = 0.21$, $P = 0.6$) of the exiting object had any effect.

A mixed ANOVA with class numerosity (1-instance/3-instances), distance (near/far) as within-participant factors and order of presentation (3-instances first /1-instance first) as between participant factor revealed a trend towards an interaction ($F(1,50) = 3,41$, $P = 0.07$). Post-hoc analysis showed that infants tended to look longer at 1-class object outcomes when they saw a 1-class object outcome first ($M_{1\text{-instance}} = 15.4$ s, $M_{3\text{-instances}} = 10.3$ s; Scheffe post-hoc, $P = 0.001$), but not when they saw a 3-class object outcome first ($M_{1\text{-instance}} = 14.1$ s, $M_{3\text{-instances}} = 13.5$ s, Scheffe post-hoc, $P = 0.5$).

Experiment 3

Participants

Twenty 12-month-old infants (M= 12.17m, range 11m 26d - 13m) participated in the experiment (9 boys). An additional 15 infants were excluded from analysis (8 fussed out, 3 exceeded maximum looking time criteria in more than two test trials, 2 looked away exactly when the ball exited the frame, 1 for experimental error).

Stimuli, Material and Procedure

Material and procedure were identical to Experiment 1, except that, in the experimental movies, the full occlusion period followed the fading-in period lasted 0.04 s, or 1 frame.

Results and Discussion

Data exceeding 2 SD were filtered out (6 out of 80 data points). Infants looked longer after seeing exiting the object which was farthest from the exit before occlusion ($F(1, 19) = 16.5$, $P < 0.001$). Fifteen out of 20 participants had such looking pattern ($P = 0.04$, binomial test). Unlike the previous experiments, the numerosity of the class of the object that exited the gumball machine had no effect ($F(1, 13) = 1.2$, $P = 0.29$).

Separate ANOVAs revealed a main effect of color (M_{Yellow}= 15.6 s, M_{Blue}= 10.8 s; $F(1, 19) = 7.4$, $P = 0.01$), but no effect of shape (M_{Cube}= 13.5 s, M_{Star}= 12.8 s; $F(1, 19) = 0.1$, $P = 0.6$). A further ANOVA with color, class numerosity and distance as within subject factors revealed that infants' preference for outcomes with yellow objects did not interact with numerosity or distance. Thus, we attribute this unexpected effect to group variability.

Mixed ANOVA with order of presentation as between subject factors and numerosity and distance as within subject factors revealed no main effect of order or interactions with the other factors.

Formal Model Description

Formally, according to our model the degree to which observers expect the final outcome, D_F, is given by how well they could have predicted that outcome from the starting conditions (D_0), any observed data ($D_{1\ldots F-1}$), and their world knowledge (**W**):

$$P(D_F \mid D_{0,\ldots,F-1}, W) . \quad (1)$$

This amounts to summing over all possible world states that the observer conceived given the starting conditions (S_k; because this probability distribution is approximated with Monte Carlo, we consider K possible world states – the main results in the paper used K=20,000, while results varying K from 1 to 512 world states are presented below),

$$\propto \sum_{k=1}^{K} P(D_F, S^k_{0,\ldots,F} \mid D_{0,\ldots,F-1}, W)$$
$$\propto \sum_{k=1}^{K} P(D_F \mid S^k_{0,\ldots,F}) P(S^k_{0,\ldots,F} \mid D_{0,\ldots,F-1}, W) \quad (2)$$

where $P(D_F \mid S^k_{0,\ldots,F}) = (D_F \approx S^k_F)$, and $\approx$ corresponds to Boolean consistency with a set of key features.

In our case,

$$P(S^k_{0,\ldots,F} \mid D_{0,\ldots,F-1}, W) \propto \prod_{t=1}^{F} P(S^k_t \mid S^k_{t-1}) P(D_{t-1} \mid S^k_{t-1}) , \quad (3)$$

thus yielding the probability of the final observed world state as:

$$P(D_F \mid D_{0,\ldots,F-1}) \propto \sum_{k=1}^{K}\left[P(D_F \mid S_F^k)\prod_{t=1}^{F}P(S_t^k \mid S_{t-1}^k)P(D_{t-1} \mid S_{t-1}^k)\right]$$

In this equation, the constraint of consistency with data, $P(D_t \mid S_t^k)$, uses the same definition of "consistency" as used in the prediction of the final outcome.

The transitions, $P(S_t^k \mid S_{t-1}^k)$, are given by constrained Brownian motion on all of the considered objects:

$$P(S_t^k \mid S_{t-1}^k) = \prod_j P(x_t^j \mid x_{t-1}^j),$$
$$P(x_t^j \mid x_{t-1}^j) = N_c(x_{t-1}^j, \sigma)$$

where x corresponds to the position of a single simulated object, and N_c is a constrained Gaussian distribution with standard deviation σ. This constrained Gaussian expresses knowledge about solidity, which prevents objects from passing through solid walls (Fig. S1). Brownian motion yields a simple formalization of spatiotemporal continuity of objects, since the expected change in an object's position is centered on zero, and is more likely to be small than large.

Additional Model Results

Unlike with the experiments we presented in the main text, quantitative model fits are not possible with much of the literature on infant cognition, where each study typically includes only one or two binary contrasts, yielding insufficient data for meaningful correlations, and too many other factors contributing to looking times vary across experiments to allow cross-experiment analyses. Nevertheless, we can demonstrate qualitative consistency between the observed binary effects and the predictions of our model. We now present the analysis of some crucial experiments about infants' cognitive abilities relative to physical and probabilistic situations, and compare them with our model's predictions.

1. Xu and Garcia *(S4)*: Relationship between samples and a population

In the main text, we have shown that our ideal observer model predicts infant behavior in a set of novel and classic experiments about infants' intuitions of the probability of a single future event obtained with different paradigms by Téglás and his collaborators *(S2)*.

However, forward simulation for prediction includes, as a simple case, simple probabilistic calculations from observed frequencies. Therefore, our model should also account for other key experiments on infants' statistical reasoning. Consider Xu & Garcia's demonstrations that infants understand the relations between samples and populations *(S4)*. When infants see a box containing mostly red and few white balls, they look longer at a sample containing mostly white balls, indicating that they are surprised to see unrepresentative samples from a previously observed population. Infants even reason from a sample to the population: after seeing a mostly white-ball sample, they look longer when the population is revealed to contain mostly red balls. Our model accounts for these results simply by simulating possible outcomes given initially ambiguous parses of the world. Indeed, even a simpler ideal observer without any model of stochastic physics would also predict these results.

In Xu & Garcia's experiments 4-8, infants saw a box with 70 balls of type A, and 5 balls of type B; then they saw samples drawn from the box (Figure S2A).

The sampled balls were either four of type A and 1 of type B, or vice versa. In two experiments, infants looked longer at an unlikely (unrepresentative) sample (1 A, 4 B; red) than a representative sample (4 A, 1 B; blue; Figure S2B). Starting with fully observed knowledge of the contents of the box, the model forms predictions about a particular sampled ping-pong ball by simulation from these starting states. A draw from the box corresponds to the object that is hypothesized to be closest to the exit. Assuming this procedure, the model can allow one to compute the proportion of simulated outcomes consistent with seeing different sample sizes ({1,1},{2,1},{3,1},{4,1}) as a function of the box contents. This proportion increases for representative samples, and decreases for unrepresentative samples (Figure S2C). Thus, the model estimates the probability of seeing a {4,1} sample as being much higher than that of seeing a {1,4} sample. This result is the exact looking time behavior found by Xu & Garcia when infants were required to draw inferences from the populations to the samples.

Conversely, Xu & Garcia also showed that infants can reason from samples to populations. In other experiments, infants saw five ping-pong balls drawn from a box with unknown contents (Figure S2D). Four balls were of one color (A), and one of another color (B). The contents of the box were then revealed to be either 70 A-balls, and 5 B-balls, or vice versa. In two experiments, infants looked longer at the box contents from which the sample was unlikely to be drawn (5 A, 70 B; Figure S2E).

The model unfolds as follows. It assumes initial uncertainty about the world state; that is, it does not know whether the box contains mostly type-A, or mostly type-B ping-pong balls. To represent this uncertainty, the model starts with a large set of possible parses representing containers with unknown proportions of type-A and type-B ping-pong balls. Parses of the world based on these possibilities are generated and updated based on consistency with the observed samples. As the model keeps generating possible states of the world, the probability of a ~{70,5} over a ~{5,70} box increases with the skew of the observed sample (Figure S2F). Thus, after a {4,1} sample, the ~{70,5} box is much more likely than a ~{5,70} box (Figure S2F). The probabilities represented in the graph correspond to a sum of the parses from {1,74} to {9,66} and {74,1} to {66,9} – to account for infants' imperfect and noisy quantity judgments of the final display.

2. Wilcox and Schweinle *(S5)* : Reasoning about slow motion

Slow-and-smooth assumptions about spatiotemporal trajectories also predict different effects of speed and timing. Thus Wilcox and Schweinle *(S5)* showed infants scenes in which an object passed behind an occluded and emerged on the opposite side either immediately, or after a delay compatible with one object traveling a continuous trajectory behind the occluder. Thus, infants saw one object disappear behind a large occluder then re-emerge from the other side either instantaneously, or after a plausible delay (Figure S3A). Infants then either saw the state of the world fully revealed, indicating that only one object was present (unambiguous; blue), or they saw another occluder behind the displaced one (ambiguous; red), thus creating ambiguity about whether or not one or two objects were present. Infants looked longer at the unambiguous single-item state when an object emerged instantaneously on the one side of the occluder after the disappearance of the object behind the opposite side of the occluder, as if a single object moved very quickly.

Like infants, the model starts with an ambiguous parse of the world, in which either one or two objects are assumed to be present. From this, it generates future worlds to predict when an object will emerge from the right side of the screen (Fig. S3B). Under a two-object parse, an object is much more likely to emerge on the right side of the screen without delay than under a one-item parse. Thus, depending on the observed time of emergence, two-object parses are much more likely if an object emerged early (Fig. S3D). Integrating these predictions over a range of time, we obtain a cross-over interaction, in which an immediate re-emergence would suggest two items, but a delayed re-emergence suggests one item. Since the ambiguous state makes both one- or two-object parses likely, the final observed state is unlikely (unexpected) only if it is (1) unambiguous (blue) and (2) involves an immediate re-emergence. Instead, a delayed re-emergence or an ambiguous final state (red) are always predictable (Fig. S3F). Integrating predictions over time, we obtain the probability of the final display given the current predictions in the immediate and delayed conditions, which can be interpreted as the analog of infants' surprise. Indeed, the models' predictions inversely match the looking times of the participants in Wilcox and Schweinle's experiments (Figure S3G).

A constrained observer using few samples.

Thus far we have described our model at a purely computational level: Which information is used at the onset? What knowledge and assumptions are brought to bear on the predictions? How is surprise evaluated with respect to the set of predicted outcomes? These computations have been cast in terms of ideal Bayesian inference which requires keeping track of potentially infinitely large hypothesis spaces, and updating them all with further observations – indeed, the predictions for the ideal observer that we describe were all computed with thousands of Monte Carlo simulations each. It may be argued that although infants exhibit behavior consistent with such an ideal observer, it would be cognitively implausible to suppose that they are doing anything so computationally intensive. In this section, we will describe how the calculations carried out by the ideal observer can be approximated with simple algorithms that might more plausibly reflect the resource-bounded computations that infants, and adults alike, use to make these predictions.

We computed expectations for our model using just a small number of conditional samples, that is, simulated worlds consistent with the observations up to a given point in time. Effectively, if the current simulated world predicts that an object should exit at time t, and an object does not exit at time t, that world extension is rejected, and the simulation continues until the sample predicts an object exit that does indeed occur, or an exit occurs that was not predicted.

There are different ways we could consider modeling a limited capacity observer by restricting the number of samples. Most consistent with the literature on boundedly rational models in adult cognition *(S6-S10)*, we could restrict the number of conditional (or posterior) samples used to calculate graded expectations. This corresponds to K in Equation 2. This is analogous to using a small number of particles in a particle filter to track the current world state, which has previously been proposed to account for human behavior *(S6, S7, S11)*. Fig. S5 shows our model predictions for the current experiments as a function of the number of conditional hypotheses (samples) that are entertained by the model, for 1, 2, 4, 8, … 512 samples.

Fig. S6 shows the correlations between model predictions with different numbers of samples, and the infant data from our three experiments. It can be seen that even with just one sample per subject and 20 subjects, as in our experiments, the rank-ordering of conditions is unchanged and the correlation remains high. With 4 samples per subject, the model predictions are practically identical to those of the ideal observer. Therefore, although throughout the paper we compare our data to the fully ideal observer (with no resource limitations), none of our conclusions depend on precise approximation of the Bayesian ideal observer with large numbers of samples.

Because of the way our samples are computed, by drawing from the prior on object motion and then rejecting samples inconsistent with features of the observed data, we could also model a resource-bounded observer by restricting the number of samples from the prior. This is a more severe restriction: with only one or a small number of samples it will often lead to finding *no* consistent conditional samples, so predictions will not be computable. It is also only cognitively natural to the extent that the rejection-based method of approximate inference we give here corresponds to the actual mechanisms of prediction in infants' minds, which is unlikely; it does not scale well to complex scenes. In contrast, there are many ways to compute approximate conditional samples, such as the particle filter methods referred to above, that do scale to complex scenes, so we find a restriction on the number of conditional samples to be more cognitively natural. Nonetheless, for completeness, Fig. S7 and S8 show model predictions for a restricted number of samples from the prior. In this setting we find that model predictions become as accurate as the ideal model once the number of samples exceeds 30.

If infants are indeed extrapolating object motion according to an internal physical model, an interesting question is how many plausible extrapolations they actually keep track of – how many samples they use. Our data do not offer the resolution necessary to distinguish between an essentially unbounded number of conditional samples and just one; however, we can tentatively speculate on the bounds within which our simulations are plausible proxies for infants' cognitive processes. Infants must at least be able to extrapolate one set of trajectories for up to two seconds with no observations -- otherwise, it would not have been possible to collect our data. In adult object tracking and memory studies, it is routinely noted that people can track four objects *(e.g., S12)*, and often more *(e.g., S13)*, arguably using a model of physical dynamics analogous to the one we propose here *(S11)*. Moreover, infants can represent arrays and sets *(S14-S16)*, and adults seem to attend to groups of objects in an ensemble *(S17)*, which would facilitate extrapolating a set of trajectories for the whole group of objects. Altogether, if infants can also represent ensembles of objects, in our experiments they would need to only represent two ensembles (one for each color category), and thus could extrapolate two to four plausible future trajectories. Of course, such proposed constraints are mere speculation, because in our experiment, even one simulated set of trajectories is adequate to account for infant behavior.

Detailed Modeling Parameters

We now present some further details about the parameters that were used to model the experiments we analyzed in this paper.

1. Current experiment (Fig. 3):

The initial observations are parsed into the circular container (7 cm in radius) with a 20 degree opening at the bottom, as well as the three identical and one unique objects. The objects that started near the exit before occlusion started at 4 cm below the center of the container, and those that started far from the exit started 4 cm above the center of the container. The objects were circular: 0.5 cm in radius. Objects moved according to constrained Brownian motion, with motion perturbations with a standard deviation of 0.5 cm. The qualitative feature that consisted of a predicted observation is an object exiting through the gap.

2. Téglás et al. *(S2)*, Gumball experiments (Fig 5A-5F)

The initial observations are parsed into the circular container (7 cm in radius) with a 20 degree opening at the bottom, and three identical and one unique objects. Objects all started in the center of the container. The objects were circular: 0.5 cm in radius. Objects moved according to constrained Brownian motion, with motion perturbations with a standard deviation of 0.5 cm. In the case in which the objects of the most numerous class were sequestered by a barrier (Figure 4D) *(S2, Experiment 2)*, the objects started 4 cm above the center of the container, while the barrier passed through the center. The qualitative feature that consisted of a predicted observation is an object exiting through the gap.

3. Téglás et al. *(S2)*, Box with exits (Fig 5G-I)

Téglás et al. (2007) also created situations where the geometry of the stimulus, rather the number of objects, represented the information about likely and unlikely outcomes. In their scenes, which are better suited to test children's expectations as measured by reaction times, a ball bounced inside a rectangular box with one hole in a wall and three in the opposite wall. After a period of visible movement, an occluder covered the box. Five-year olds had to press a button when they saw the ball exiting the box.

To test the model's prediction for these scenes, we set the initial scene as a box 15 cm tall, 9 cm wide. Its left side contained one 2 cm exit, while the right side contained three 2 cm exits, equally spaced. The ball started in the center of the box, and moved with constrained Brownian motion with a standard deviation of 0.5 cm. The qualitative feature was an emergence of the ball on either the left or right side.

4. Kellman and Spelke *(S18)* (Fig. 5A-C)

The two objects behind an occluder were ambiguously parsed into either independent objects (moved by independent Brownian motion) or conjoined objects (shared a source of movement perturbations). Brownian motion had a standard deviation of 0.5 cm. The qualitative feature evaluated at each time step was a match between the observed direction of motion of each bar end, and the predicted direction of motion. Predicting the correct direction of motion for two independent perturbations is less likely than for one shared perturbation.

5. Aguiar and Baillergeon *(S19)* (Fig 5D-F)

The model started with an unambiguous parse into one or two objects, and one or two screens. The left and right edge of the screens are at -2 and +2 cm away from center; when a gap is present it occupies the space between -0.5 and 0.5 cm from center. Objects move with a standard deviation of 0.5 cm. The qualitative feature detected is an object emerging on the right side after disappearing behind the left edge without appearing anywhere else (e.g., again on the left side or between the two screens when a gap is present).

6. Xu and Garcia *(S4)* (Fig S2):

From population to sample: 75 objects (70 of one type, 5 of another) are observed in a container with a small hole. Five objects are removed from the container depending on their proximity to an exit hole. The qualitative features matched to the observation is whether the set of objects contains 4 common objects and 1 uncommon object, or vice versa.

From sample to population: The model starts with an ambiguous parse – an occluded box contains either 70 objects of type *A* and 5 objects of type *B*, or vice versa. Then samples are drawn from the box as described above. The qualitative feature matched is an observation of 4 type *A* objects and 1 type *B* object. The contents of the box are then revealed.

7. Wilcox and Schweinle *(S5)* (Fig S3):

An object is observed to disappear behind the left side of a screen covering -2 to +2 cm of the display. The display parse is ambiguous: either one object is present (the one that just disappeared), or two objects are present (another behind the right edge of the screen). The qualitative feature assessed is the emergence of an object on the right edge of the screen, as a function of time. This yields an inference about the number of objects present in the display (Fig S3E). The screen is dropped to reveal either exactly one object, or one object plus (perhaps) another behind a screen. The probability of this ambiguous final display is computed by integrating out the number of objects, since either one or two object interpretations are equally consistent with the ambiguous display. However, only the one object interpretation is consistent with the unambiguous one-object display. Thus, assuming our hypothesis about the relationship between probability and surprise, the model is "surprised" to see an unambiguous one object display after a brief occlusion period because such an occlusion period makes a two-object interpretation much more likely.

8. Xu and Carey *(S20)* (Fig S4):

The structure of the experiment was first introduced by Spelke & Kestenbaum *(S21)* with an habituation paradigm. Here we model Experiment 1 of Xu and Carey *(S20)*, where the same results have been obtained without habituation. Two occluding screens are visible. In the discontinuous condition, one object is seen to emerge on the left side of the left occluder, then an object is seen to emerge on the right side of the right occluder. In the continuous condition, an objet emerges on the left side, then an object is seen in the middle, and then an object emerges on the right side. When the two screens are dropped, infants see either one or two objects behind these occluders – in the discontinuous condition, they are more surprised to see one object; in the continuous condition, they are more surprised to see two objects. To model these results, we add a further solidity constraint – that objects cannot pass through one another – and model the motion dynamics in three, rather than just two, dimensions. Objects are spheres with radius 0.5. They move according to Brownian motion with a standard deviation of 0.25. The occluders are 1.5 units wide, and the middle gap is 1 unit wide. The key features for the model are defined as sequences of observations: For the discontinuous condition (1) no objects are visible, (2) an object is visible on the left, (3) no objects, (2) then an object is visible on the right, followed (3) by no object being visible. For the continuous condition: (1) no objects, (2) an object on the left, (3) no objects, (4) an object in the middle, (5) no objects, (6) an object on the right, (7) no objects. The model starts with an ambiguous parse of one or two objects, and forward-simulates trajectories according to constrained Brownian motion further constrained by these sequential key features. We compute the probability of the observed final state (one or two objects) by tabulating the proportion of one and two object parses that could account for the entire sequence of observations. Because a single object is unlikely to "jump" across the visual gap, the discontinuous condition favors a two-object parse. However, because a two-object parse is likely to yield additional, unpredicted observations in the continuous case, the one-object parse ends up more likely in that condition.

References

S1. J. D. Cohen, B. MacWhinney, M. Flatt, J. Provost, *Behav. Res. Meth. Ins. C.* **25**, 257 (1993).

S2. E. Téglás, V. Girotto, M. Gonzalez, L. L. Bonatti, *Proc. Natl. Acad. Sci. U. S. A.* **104**, 19156 (2007).

S3. J. Unturbe, J. Corominas, *Neuropsychology* **21**, 621 (2007).

S4. F. Xu, V. Garcia, *Proc. Natl. Acad. Sci. U. S. A.* **105**, 5012 (2008).

S5. T. Wilcox, A. Schweinle, *Infant Behav. Dev.* **26**, 253 (2003).

S6. R. Levy, F. Reali, T. L. Griffiths, *Proceedings of NIPS* **21**, 937 (2009).

S7. S. D. Brown, M. Steyvers, *Cogn. Psychol.* **58**, 49 (2009).

S8. E. Vul, N. D. Goodman, T. L. Griffiths, J. B. Tenenbaum, *Proceedings of the 31 Annual Meeting of the Cognitive Science Society* 148 (2009).

S9. L. Shi, T. L. Griffiths, N. H. Feldman, A. N. Sanborn, *Psychonomic Bulletin & Review* **17**, 443 (2010).

S10. A. N. Sanborn, T. L. Griffiths, D. J. Navarro, *Psychol. Rev.* **117**, 1144 (2010).

S11. E. Vul, M. Frank, G. Alvarez, J. Tenenbaum, *Advances in Neural Information Processing Systems* **22**, 1955 (2009).

S12. B. J. Scholl, Z. W. Pylyshyn, *Cogn. Psychol.* **38**, 259 (1999).

S13. G. A. Alvarez, S. L. Franconeri, *J Vis* **7**, 14.1 (2007).

S14. L. Feigenson, J. Halberda, *Cognition* **91**, 173 (2004).

S15. L. Feigenson, S. Carey, M. Hauser, *Psychol. Sc.* **13**, 150 (2002).

S16. L. Feigenson, S. Carey, *Cognition* **97**, 295 (2005).

S17. G. A. Alvarez, *Trends Cogn. Sci.* **15**, 122 (2011).

S18. P. J. Kellman, E. S. Spelke, *Cogn. Psychol.* **15**, 483 (1983).

S19. A. Aguiar, R. Baillargeon, *Cogn. Psychol.* **39**, 116 (1999).

S20. F. Xu, S. Carey, *Cogn. Psychol.* **30**, 111 (1996).

S21. E. S. Spelke, R. Kestenbaum, *Psychologie Française* **31**, 67 (1986).

Figures and Figure Captions

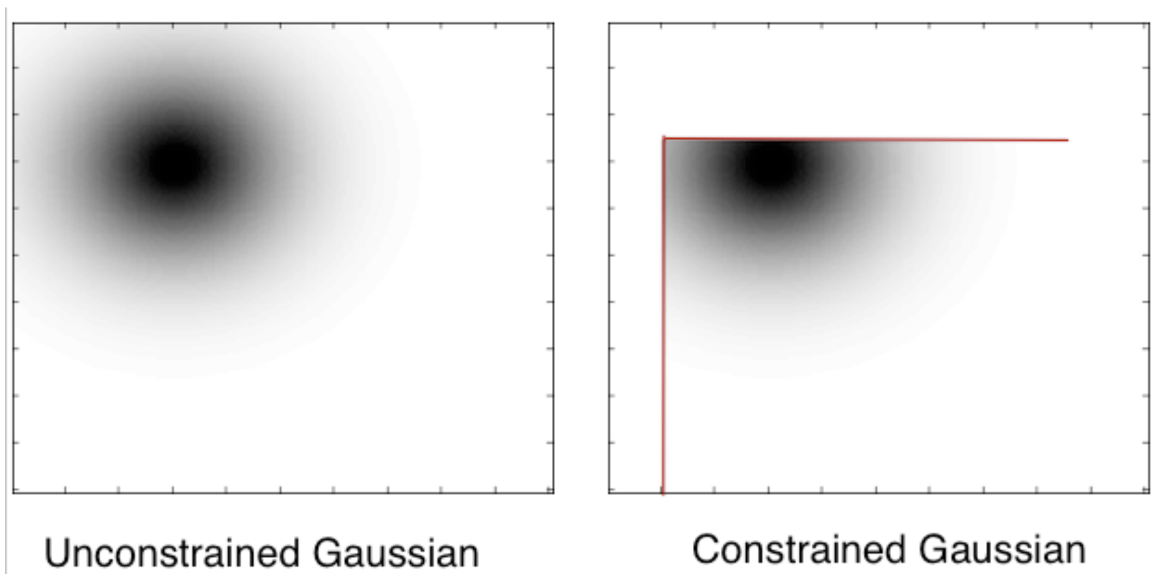

Fig S1: Demonstration of constrained and unconstrained Brownian motion. Left panel: Brownian motion predicts Gaussian transitions from *t* to *t+1*. Right panel: Constrained Brownian motion restricts the transitions to those that are allowed given the solid barriers present in the current display.

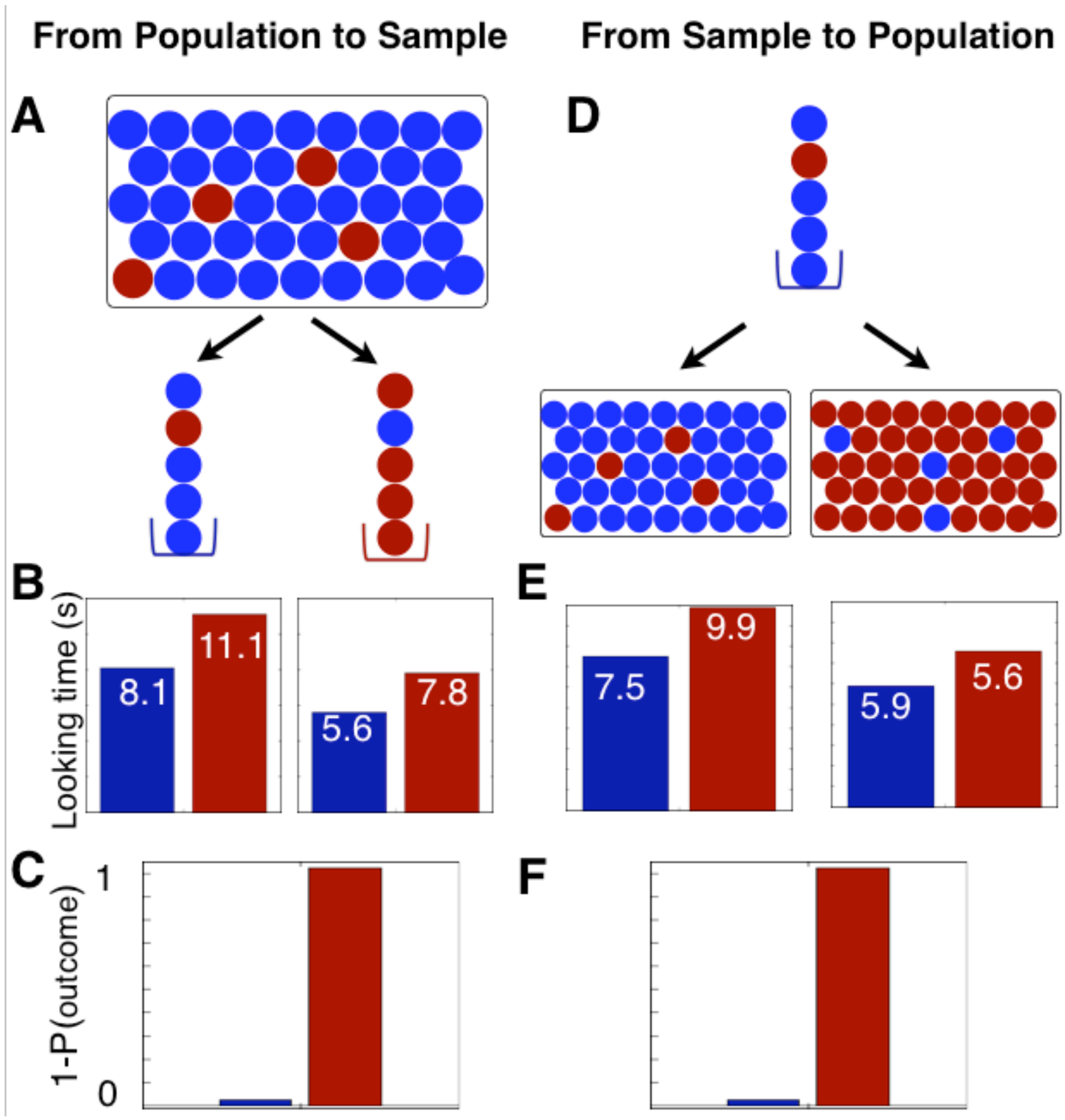

Fig S2. **(A)** Schematic representation of Xu & Garcia *(S4)* "From Population to Sample" Experiments: infants are shown a container with a 70:5 distribution of balls (e.g., blue and red), then they see a sample of (e.g.) 4 blue, 1 red, or vice versa. **(B)** In these scenarios infants look longer at the improbable sample. **(C)** Similarly, the improbable sample is a less likely outcome for our model when forward simulating from the starting conditions. **(D)** Schematic representation of

Xu and Garcia's *(S4)* "From sample to population" experiments: Infants see a sample of (e.g.) 4 blue and 1 red ball drawn from an unobserved container, then the container is revealed to either contain 70 blue and 5 red, or 5 blue and 70 red balls. **(E)** Infants look longer at the container with a population that does not match the sample. **(F)** For our model, the initial parse is ambiguous – there is an unknown proportion of red and blue balls in the container, as the sample is observed, only those parses with more blue than red balls remain; thus, our model is "surprised" to see a container with a majority of red balls (see text for details).

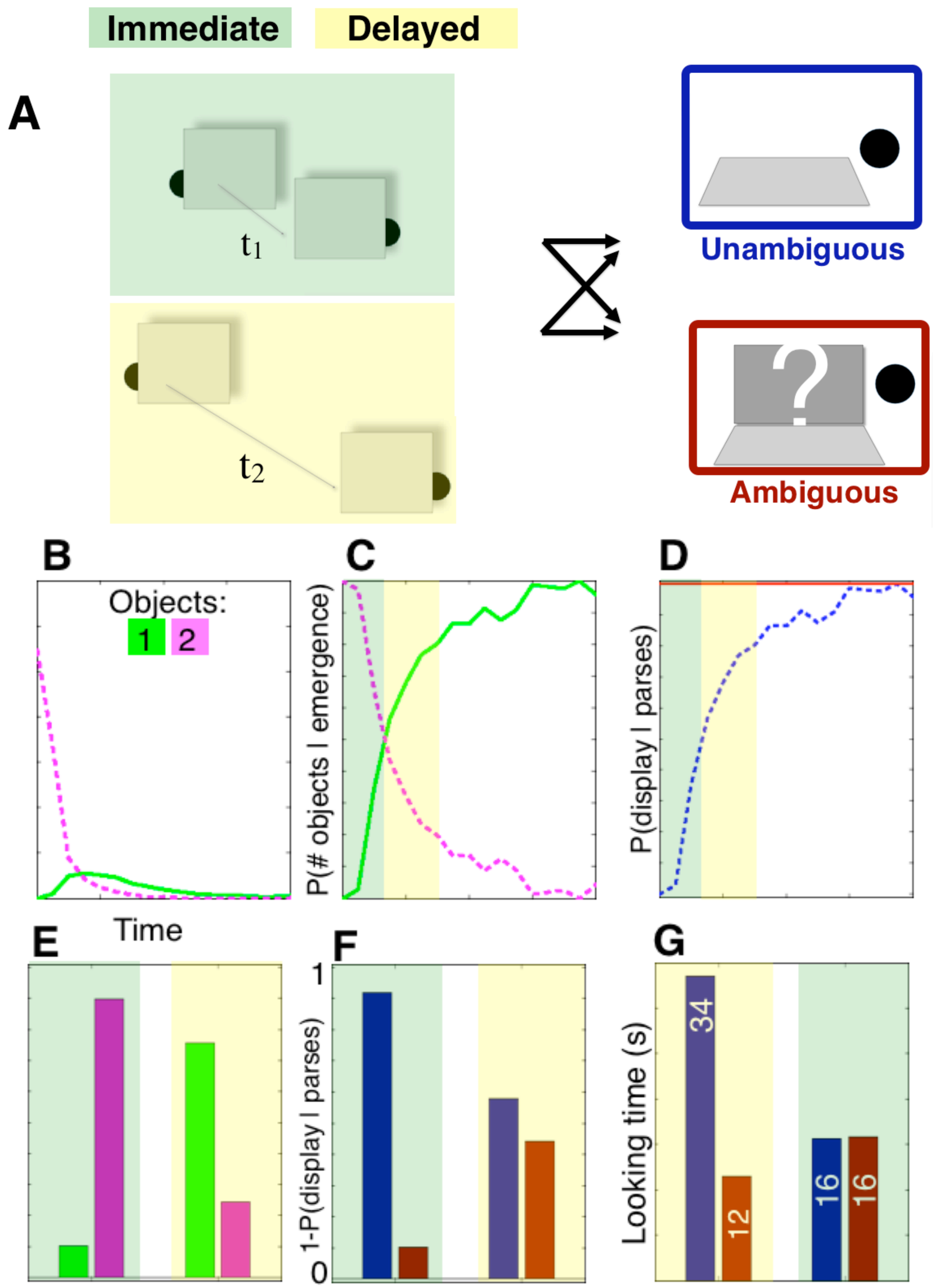

Fig. S3: Wilcox and Schweinle *(S5)* results. **(A)** Wilcox and Schweinle showed infants a display similar to Baillergeon (20), in which one ball was seen disappearing behind an occluder and a ball was seen emerging from the other side after either a short, or a long delay. Then the occluder dropped, revealing either an unambiguous display with only one ball, or an ambiguous display with another occluder that may potentially hide another ball. **(B)** According to our model, a ball is likely to emerge from the opposite side after a short delay only if there were two balls in the screen at the onset. **(C)** Thus, if the second ball emerges earlier, the display is more consistent with a two-ball parse; but if the second ball emerges later, the display is more consistent with a one-ball parse. **(D)** Thus, the probability of the unambiguous final display showing only one ball increases with the latency of the emergence of the second ball; however, the ambiguous parse is consistent with both one- and two- ball interpretations, and therefore is equally well predicted regardless of when the final ball emerges. Putting this into convenient bar-graphs: **(E)** When the second ball emerges immediately, the display is much more consistent with a two-ball parse, when the second ball emerges with a delay, the display is more consistent with a one-ball parse. **(F)** However, the ambiguous display is consistent with both parses, therefore only the unambiguous (blue), immediate emergence condition presents infants with a display that is inconsistent with their predictions. **(G)** This is precisely the display at which infants look longer.

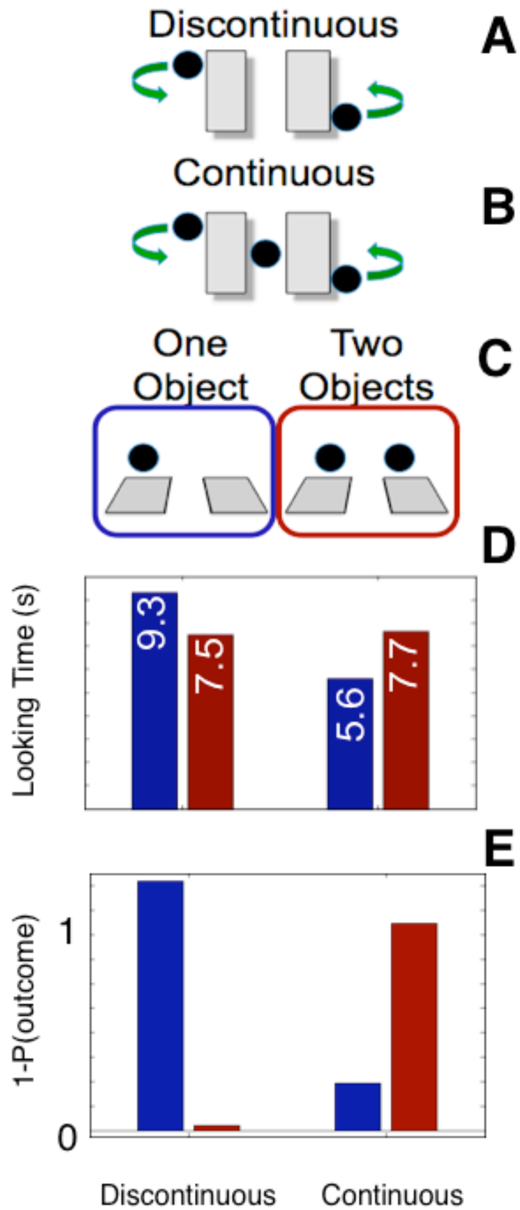

Fig S4. Xu and Carey *(S20)* results (Experiment 1) and modeling comparison. Infants saw two screens. In the discontinuous condition **(A)**, an object emerged on the left side, then the right side, but not in the middle. In the continuous condition **(B)**, an object emerged on the left side, then the middle, and then the right side. Following this display, the screens dropped revealing either one or two objects **(C)**. **(D)** In the discontinuous condition, infants looked longer when the final display revealed one object (blue) rather than two objects (red); in contrast, in the continuous condition, infants looked longer at two objects than one. **(E)** Our model predicts a similar result: due to the implausibility of a jump of one object across the two screens without showing up in the middle, the discontinuous condition favors the two object parse, thus a final display of one object is more surprising. In the continuous condition; however, the two object parse is less likely, as two objects are likely to yield many additional appearances of objects that are not observed, thus a one-object parse ends up being more likely (although not as much as the two-object parse is more likely in the discontinuous condition).

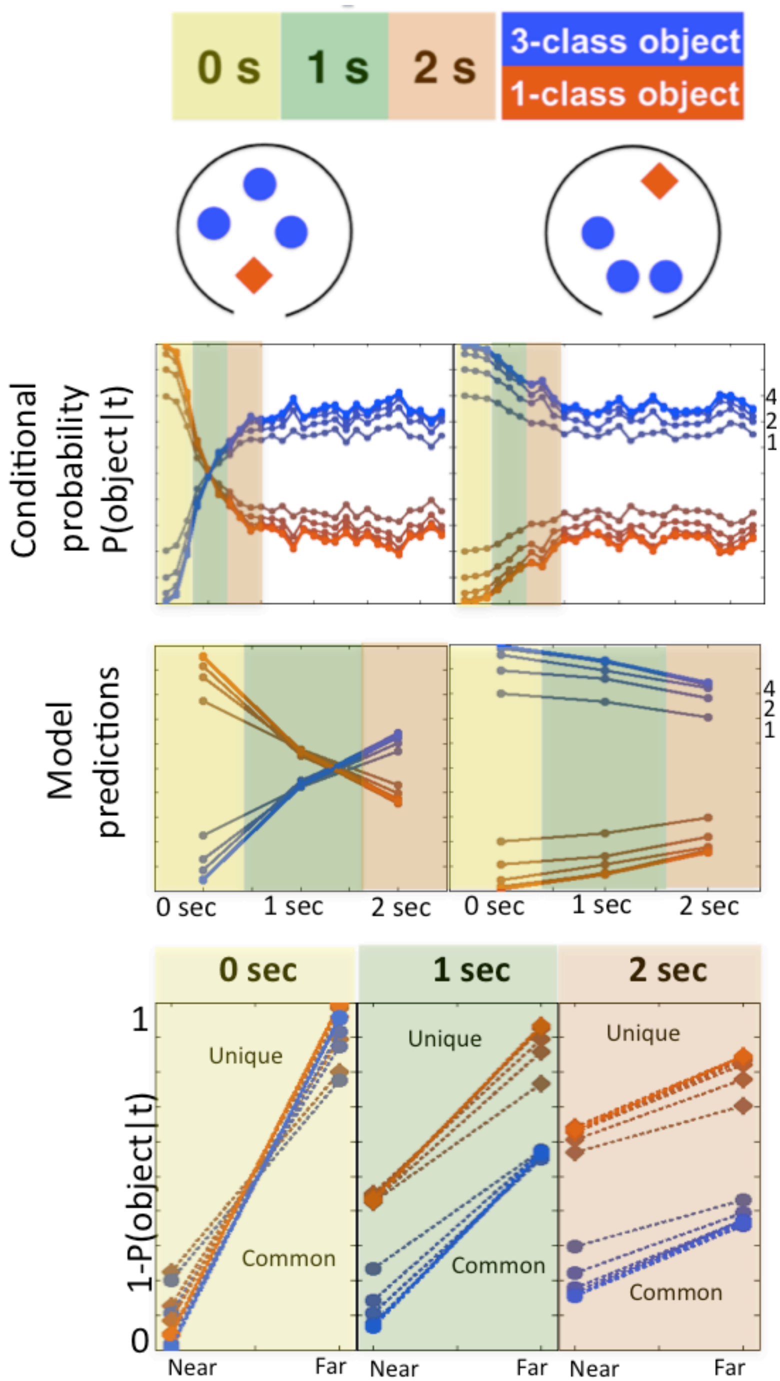

Fig S5. Model predictions using smaller sets of simulations (1, 2, 4, 8, …, 512). Each line here represents the predictions for an average over infants, each of which is using a small number of possible worlds to predict future outcomes. This figure follows panels C, D, and E of Figure 3 in the main text. Here, multiple lines represent a different number of “samples” per infant (1, 2, 4, 8, … 512). Less vibrant colors represent fewer samples and tend towards undifferentiated (50-50) predictions. However, as few as 2 or 4 samples yield the same qualitative behavior we found in our experiments.

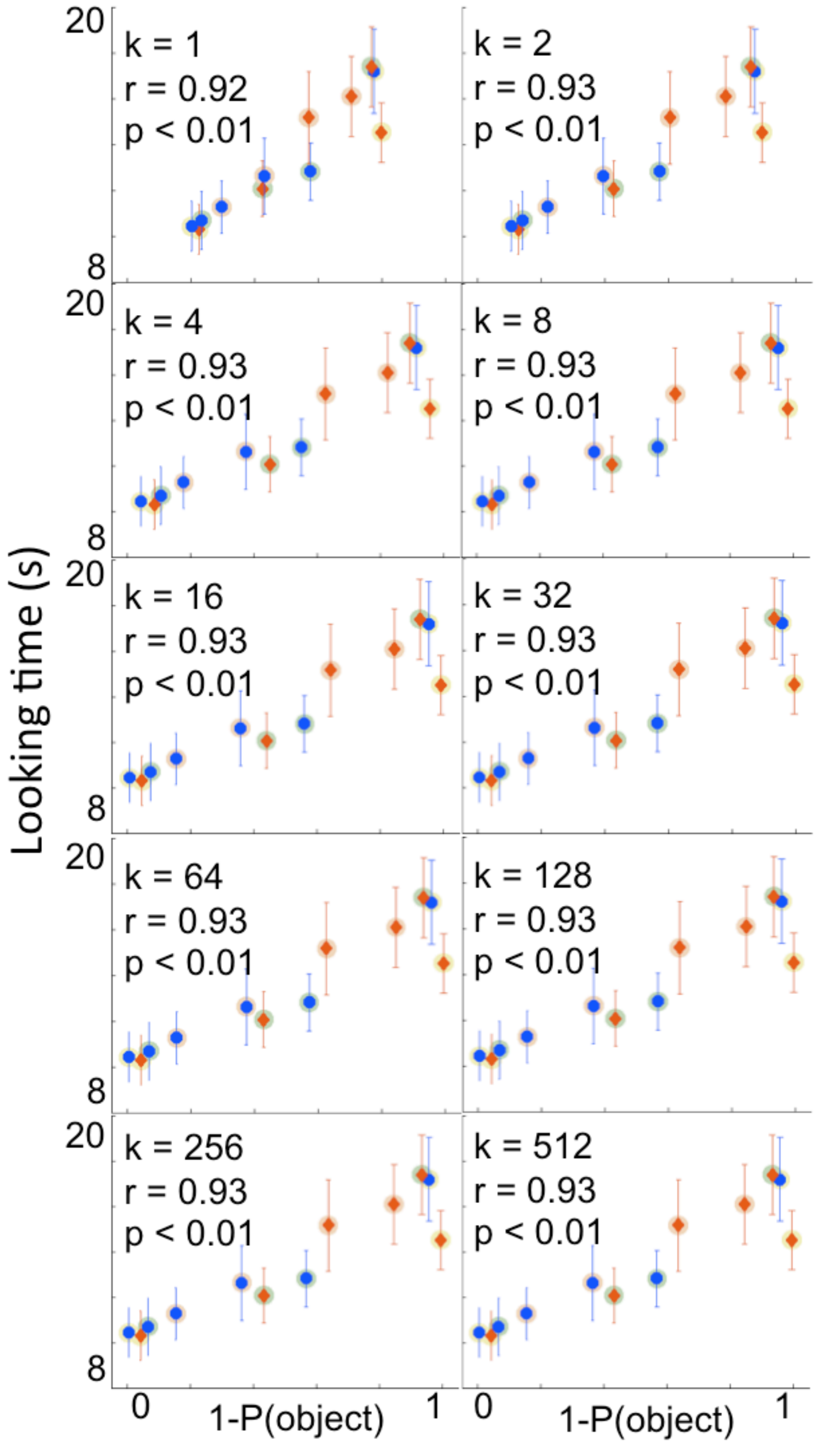

Fig S6. Correlation between model predictions given a limited number of samples (k) – like Figure S2, but for the constrained observer.

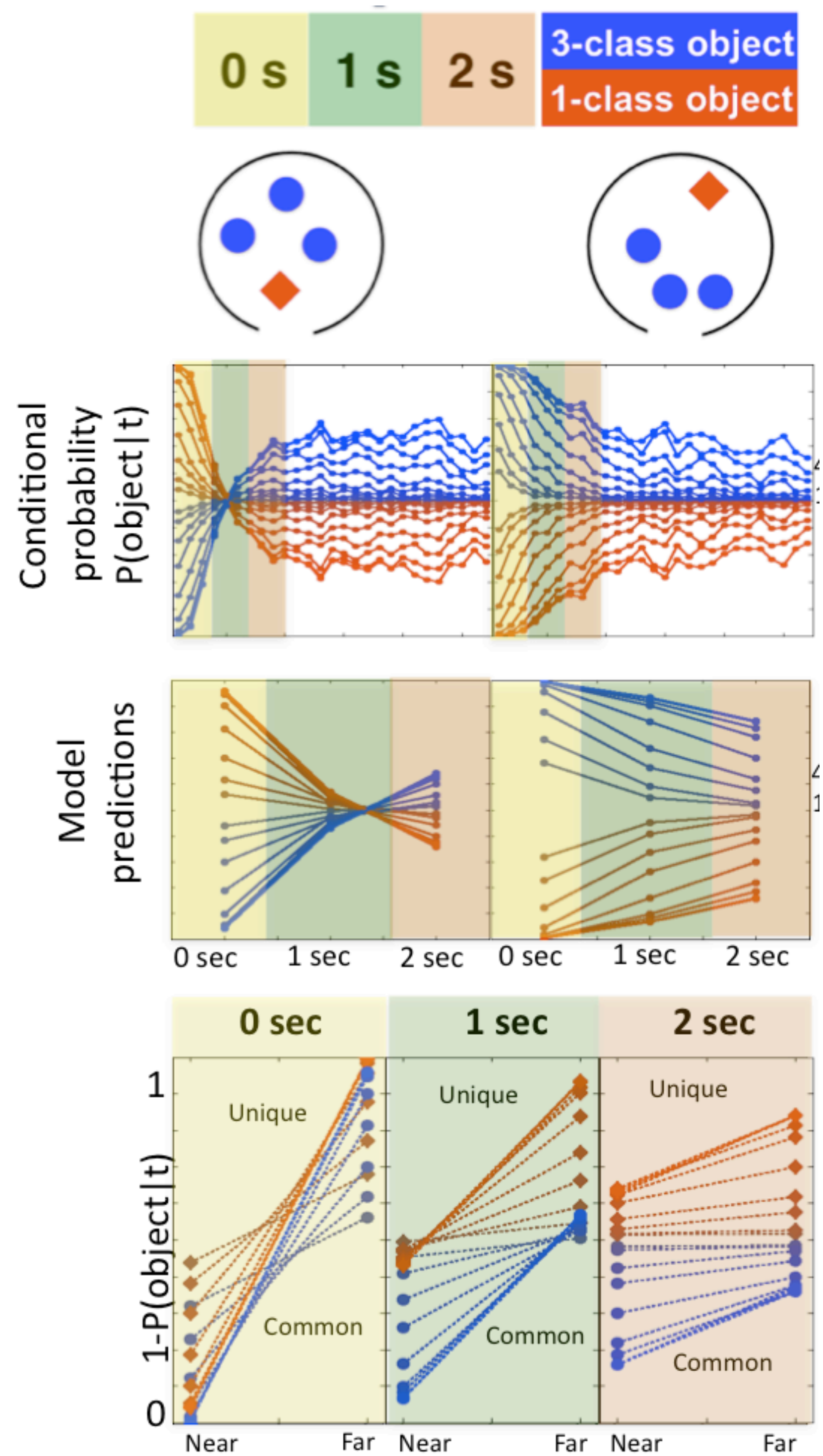

Fig S7. Model predictions using smaller sets of *prior* simulations (1, 2, 4, 8, …, 512) (not posterior, as in Fig. S5; see text). Each line here represents the predictions for an average over infants, each of which is using a small number of possible worlds to predict future outcomes, unlike figure S5, these are worlds sampled from the prior, and thus might not be rejected. This means that each sample has a high probability of not contributing to predictions. This figure follows panels C, D, and E of Figure 3 in the main text. Here, multiple lines represent a different number of "samples" per infant (1, 2, 4, 8, … 512). Less vibrant colors represent fewer samples and tend towards undifferentiated (50-50) predictions.

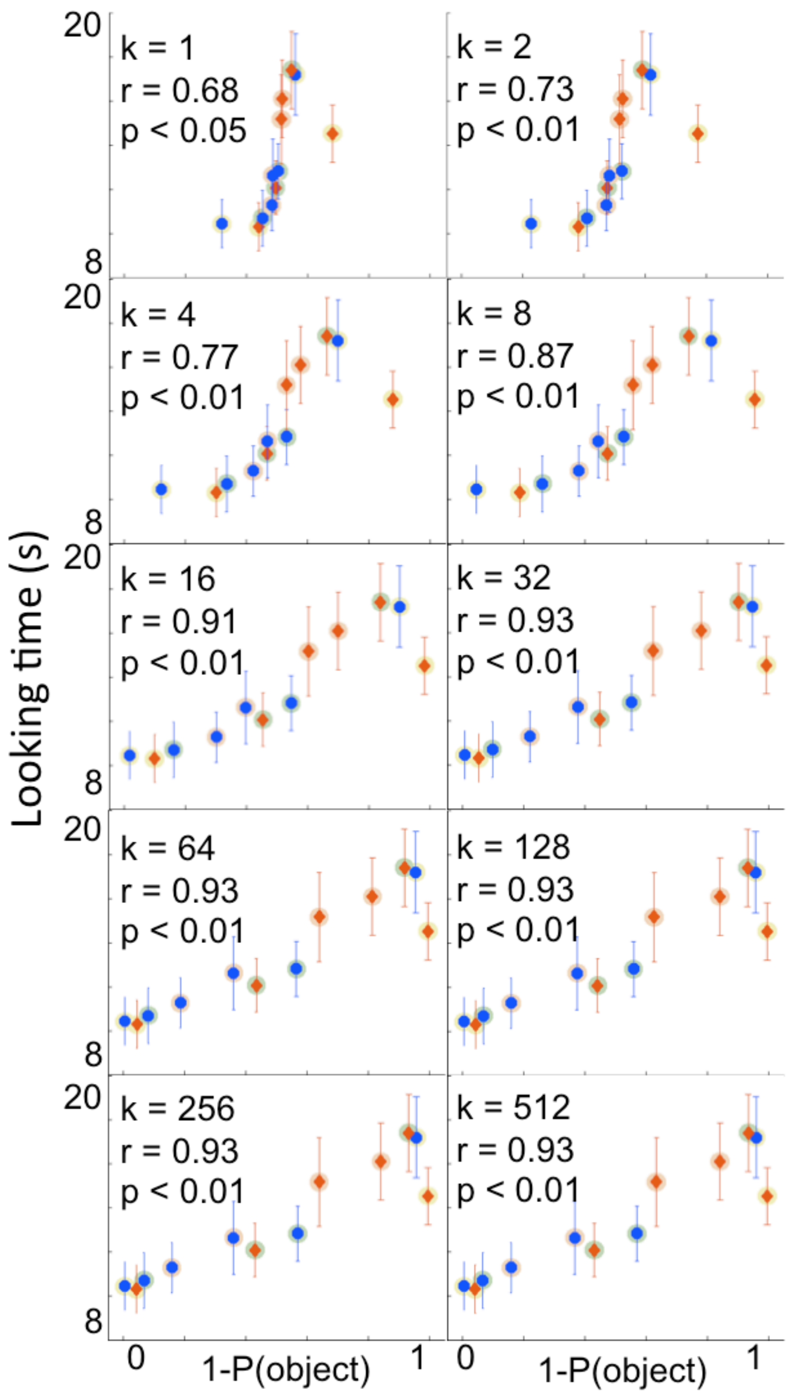

Fig S8. Correlation between model predictions given a limited number of samples (k) – like Figure S2 and S6, but for the constrained observer using a limited number of samples from the prior (not posterior as in S6).