www.sciencemag.org/content/359/6381/1263/suppl/DC1

Supplementary Materials for

Precursors of logical reasoning in preverbal human infants

Nicoló Cesana-Arlotti,* Ana Martín,* Ernő Téglás, Liza Vorobyova, Ryszard Cetnarski, Luca L. Bonatti*

*Corresponding author. Email: nicolocesanaarlotti@gmail.com (N.C.-A.); ana.martin@upf.edu (A.M.); lucabonatti@mac.com (L.L.B.)

Published 15 March 2018, *Science* **359**, 1263 (2018)
DOI: 10.1126/science.aao3539

This PDF file includes:

Other Supplementary Material for this manuscript includes the following:
(available at www.sciencemag.org/content/359/6381/1263/suppl/DC1)

MATERIALS AND METHODS

Experiment 1

In Experiment 1, 19 month-olds watched scenes in which the identification of the object inside the cup always required disjunctive syllogism. Such inference could occur when the occluder lowered, revealing the object behind it, and hence allowing infants to deduce the hidden content of the cup (Potential Deduction phase; Fig. 1).

Participants

Twenty-four healthy full-term 19-month-old infants (M= 19m 1d, range 18m 2d - 19m 23d) were retained for the analyses (8 girls). An additional 21 infants were tested but not retained due to either fussiness (6), crying (5), caretakers' interaction (2), equipment failure (1), experimenter error (6) or insufficient valid samples (1). The participants' caregivers were recruited at Barcelona maternities, and were later contacted by telephone. Participants were given a T-shirt and a certificate of attendance for their participation.

Materials

The stimuli consisted of 28 movies, prepared with the software Keynote 5.0 and exported as QuickTime movies (at 60 fps, 1024 × 768 pixels, MPEG-4 Video compression). There were four familiarization movies and 24 test movies. The aim of the familiarization movies was to acquaint infants with the events and the objects appearing during the experiment, so that novelty reactions extraneous to the experimental conditions tested could not affect the test phase.[1] Across the experiments, familiarization movies had to change in order to adapt to the tested population and the specific aspects of the experiments' design. The reader is invited to examine Table S1 for a summary of their structure and function in each experiment.

Test movies began with a still empty gray cup in the right-hand region of the scene. After approximately 0.5s, two objects entered the left part of the scene from above, in succession. The objects had different colors and shapes and belonged to different categories, but had an identical upper part. Both objects moved in their characteristic way (always accompanied by sounds), in succession, for 2s, before a dark gray occluder emerged and completely covered them. At this moment, the cup flew behind the occluder, entering at its exact center so that no clue was provided as to which object would be taken, and re-emerged with one object inside, landing in the right-hand region of the scene. The upper part of the object was always visible inside the cup, but the object

---

[1] In Familiarization movie 1, two objects (a white bear and a red car) are on the scene, then covered by an occluder and then revealed once the occluder is removed. In Familiarization movies 2 and 3, only one of the two objects was on the scene. Then it was covered by the occluder and was finally revealed once the occluder was removed. At this point the cup scooped the object, moving it to a different location and eventually revealing it again. In Familiarization movie 4, all the six objects that would appear in the test movies lay side by side and performed some movement, in succession.

identity could not be determined. While inside, the object made a pulsating movement, neutral with respect to potential clues to its identity, for 2s, accompanied by a neutral sound. Then the occluder slid down and revealed the second object, which began to make its characteristic movements in place. We call this phase "the Potential Deduction phase". After this phase, the object partially occluded inside the cup pulsed again, the scene remained motionless for about 1 s, and then the visible object exited the scene. Finally, the cup revealed its content. In half of the test movies, the object revealed inside the cup was the second object of the pair (Consistent outcome). In the other half of the test movies, it was identical to the object that had left the scene (Inconsistent outcome). The total length of each test movie was 20.35 s (see Movies S1 and S2).

The twenty-four test movies consisted of the combinations of outcome type (Consistent/Inconsistent), pair of objects (Doll-Umbrella/Flower-Dinosaur/Star-Triangle; fig. S1, A-C), initial distance of the scooped object from the cup (Near/Far), and identity of the object scooped by the cup.

The movies were played on a 24-inch screen by the software PsyScope X (http://psy.cns.sissa.it/), which controlled the experiment. The experiment ran on an Apple Mac Pro Quad Core 2.8 computer. The screen covered a 35 × 26.5-cm area. The movies were accompanied by sounds associated with the elements of the scene. Sounds played contingent with the events in the movies. For example, the occluder rising and falling was associated with a specific sound; the movements of the cup were associated with another specific sound; and so on (see for example Movie S1). A camera hidden behind the screen digitally recorded the 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

The experiment took place in a sound-attenuated room with dimmed lights. Participants seated on their caregivers' laps, at about 60 cm distance from the monitor. The caregivers wore opaque glasses that prevented them from seeing the stimuli. They were instructed to keep the child seated on their laps and not to interact with them. They were invited to hold infants at their hips with both hands and let them move freely, but not to let infants abandon their initial posture (facing the screen) for more than 5 s. After this period, they had to gently turn them back towards the center. In this way, reorientation towards the screen did not interfere with timeout triggering. The experimenter, who was blind to the experimental conditions, monitored infants' behavior outside the room, from a separate screen via an infrared camera. The presentation of the stimuli was participant-controlled: the movies paused or played according to whether infants looked at the screen. This procedure ensures that each infant sees each movie entirely.

The experiment consisted of a familiarization phase and a test phase. In the familiarization phase, participants saw four movies (in two orders 1-2-3-4, and 1-3-2-4, counterbalanced across participants). Familiarization movies were designed to introduce participants to the subcomponents of the scenes (the objects in the scenes, and the events

of occlusion and containment) without ever showing the exact scene to be tested in the test phase. As a result, during familiarization, participants never saw which object was inside the cup after the scooping events (for a summary of the familiarization structure and function, see Table S1).

In the test phase, participants saw six test movies. The type of outcome (Consistent/ Inconsistent) and the initial distance of the scooped object from the cup (Near/Far) were administered in two orders (CIICCI/ICCIIC and NFFNNF/FNNFFN) counterbalanced across participants. Each participant was tested twice with each object pair. The order of presentation of the pairs was (123123), with the initial pair rotating across participants.

At the end of each movie, the last frame remained still and looking time monitoring began. The experimenter monitored looking time by pressing the mouse when a participant was directing their gaze towards the screen and releasing it when the participant looked away. A trial ended when participants looked away for at least 2.5 s or when they looked at the screen for a total of 35 s. These criteria were selected to allow for possible experimental error during online tracking. Data were then coded offline frame by frame. In offline coding, a trial was considered to end when a participant looked away for 2 s or cumulative looking time exceeded 30 s.

Results

Here we report the analyses of looking time at the outcomes, following VOE methodology. We report the analyses of oculomotor data in a separate section comprehensive of all infants' experiments (see "Analysis of Infants' Oculomotor Responses").

Participants were excluded from the analyses if they had less than one valid trial per outcome type (consistent or inconsistent) or if they had a cumulative looking time of 30 s in more than half of the trials. A trial was considered invalid, and not included in the analyses in any of the following cases: the caretaker interacted (verbally or otherwise) with the infant without complying with the instructions; the experimenter erroneously triggered the end of a trial before a 2 s look-away period; the participant looked at the outcome for less than 1 cumulative s; or looking time exceed 2.5 absolute deviations from the median, computed per condition. According to these criteria, 11% of the inconsistent trials and 10% of the consistent trials were excluded from analysis.

According to the design, when the cup revealed the object, each of the 6 trials was meant to reveal a distinct object. Because of experimenter's error, one object appeared in two trials instead. However, because the repeated object was latin-squared across participants, we did not exclude the repeated trials from the analyses.

Preliminary analyses of variance (ANOVAs) to assess the role of the cup's initial distance of the scooped object (Near vs. Far) revealed no effect. Thus, we collapsed the cup's initial distance from the scooped object in successive analyses. In all future analyses, we will not report preliminary analyses unless they reveal an effect. An ANOVA with object pair as an independent variable (Star-Triangle/Dinosaur-Flower/ Doll-Umbrella) revealed different looking times for the couples. Hence, we maintained this factor in the main analysis. The two factor ANOVA with outcome type and object pair as

a within-participant factor confirmed the main effect of object pair ($M_{Star\text{-}Triangle}$ = 9 s, $M_{Dinosaur\text{-}Flower}$ = 7.2 s, $M_{Doll\text{-}Umbrella}$ = 10.6 s; $F(2, 94) = 4.1$, $P = 0.02$). More importantly, it detected a main effect of outcome type ($M_{Consistent}$ = 7.7 s, $M_{Inconsistent}$ = 10.3 s; $F(1, 94) = 5.95$, $P = 0.017$). Crucially no interaction occurred between pairs and outcome type, showing that infants looked longer at inconsistent outcomes regardless of the pair of objects which were on display. Thus, we conclude that nineteen-month-olds looked longer at the inconsistent than at the consistent outcome. A non-parametric analysis on the average looking times for the consistent/inconsistent outcomes confirmed this result (Wilcoxon matched-pairs Signed-rank $z = -2.28$, $P < 0.022$, two-tailed).

Experiment 2

Conceptually, Experiment 2 was exactly as Experiment 1, modulo the modifications suggested by the younger age of the infants tested.

Participants

Twenty-four healthy full-term 12-month-old infants (M= 12m 9d, range 11m 27d - 12m 30d) were retained for the analyses (11 girls). An additional thirteen infants were tested but not retained due to fussiness (6), caretakers' interaction (1), equipment failure (3), experimenter error (2) or insufficient valid samples (2). Participants were recruited as in Experiment 1.

Materials

The materials of Experiment 2 were identical to those of Experiment 1 except for the changes specified below. Those changes were guided by pilot studies which suggested adaptations of the material and procedure to a younger age, as well as the results of Experiment 1, which found baseline differences in looking times for the couples we used. Their overall aim was to reduce the memory demands of the task to a more tolerable level for this age, and to attract participants' attention to the identical top part of the objects. First, we increased the number of familiarization movies (6) but never presented participants with the test objects, so that they would appear more interesting at test[2]. Second, for the test phase we eliminated the pairs that were less looked at in Experiment

---

[2] In familiarization movie 1 and 2, one object (a red puppet in movie 1 and a red star in movie 2) was on the screen, then was covered by the occluder and then revealed once the occluder was removed. The cup scooped the object, it moved it to a different location and eventually revealed it again. In familiarization movies 3 and 4, only one of the objects was on the scene (the red puppet in movie 3 and the red star in movie 4), then covered by the occluder. Afterward, the cup scooped the object behind the occluder, it moved it to a different location and eventually revealed it. In familiarization movies 5 and 6, two objects (the red puppet and the red star) were on scene, then covered by the occluder. Afterward, the cup scooped one of the objects behind the occluder, and it moved it to a different location. Eventually, the occluder was removed, and an object behind it was found, while the other object remained hidden inside the cup.

1, retaining only the Doll/Umbrella pair, and we generated a novel animate/inanimate pair, composed of a novel Doll and a Flower (fig. S1, A and D). Third, after the two objects entered the scene, the occluder raised only partially, leaving the common upper part of the object visible, then it lowered, and finally it fully raised. The occluder remained stationary in this state for 1 s and then slid down and revealed the two objects. The objects remained entirely visible for about 1s. Then the occluder raised again up to the objects' identical part (leaving it visible), it remained stationary for 1 s and eventually raised to cover them completely. Afterward, the movie continued exactly as in the test movies of Experiment 1. Fourth, we reduced the overall occlusion length from the 13.75s in Experiment 1 to about 5.8 s, so as to reduce possible memory decay caused by long occlusions. The total length of the test movies was 17.51s.

The resulting test movies (16) were obtained by the combinations of outcome type (Consistent/Inconsistent), the pair of objects (Doll1-Umbrella/Doll2-Flower), object scooped by the cup (Doll1/Umbrella/Doll2/Flower) and initial distance of the scooped object from the cup (Near/Far). Movies played on a Tobii T60XL eye-tracker (http://www.tobii.com) run by the software PsyScope X, and covered an area of 28.3 x 22.4 cm. In all experiments analyzing looking times, data were analyzed with the DataDesk 7 Statistical Analysis software (https://datadescription.com).

Procedure

The procedure of Experiment 2 was identical to the procedure of Experiment 1 except for the following changes. First, an eye-tracker recorded participants' looking behavior; the number of familiarization trials increased to 6, and their content changed as explained above; the test trials were reduced to 4. Pilot experiments suggested that these changes were beneficial for 12-month-old infants to complete the task. Crucially, none of these changes modified the logical nature of the scenes. Before the experiment, participants underwent a 5-point calibration procedure, programmed in PsyScope X. During calibration, while colorful images were played on screen, the distance from the screen (optimally, about 650 mm) and their eye height with respect to the screen (its upper half) were adjusted. Then the calibration points played in succession, and the procedure was repeated until five valid points were obtained. Once the calibration was complete, the familiarization phase began.

The six familiarization movies played in two orders (1-2-3-4-5-6, and 2-1-4-3-6-5) counterbalanced across participants. For test movies, orders of presentation (CICI/ICIC), initial distance of the scooped object from the cup (NFFN/FNNF), and object pairs (1122/2211) were counterbalanced across participants; furthermore, object revealed by the cup was different for each trial. The eye-tracker data were used to automatically calculate participants' looking time. The criteria for trial end were as in Experiment 1.

Results

Rejection criteria were identical to those of Experiment 1. Seven percent of the inconsistent and 11% of the consistent trials were excluded by the median filter.

An ANOVA with outcome type as a within-participants factor, participant as a random variable and mean looking time as the dependent variable revealed that 12-

month-olds looked longer at the inconsistent than at the consistent outcome ($M_{Consistent}$ = 6.2 s, $M_{Inconsistent}$ = 7.6 s; $F(1, 23) = 5.19$, $P = 0.032$). A non-parametric analysis revealed a strong trend consistent with this result (Wilcoxon matched-pairs Signed-rank, $z = -1.94$, $P = 0.052$, two-tailed).

Experiment 3

In Experiment 3, as in Experiment 1, 19 month-olds saw movies in which the identification of the object inside the cup always required disjunctive elimination. The main difference with Experiment 1 was that here the occluder always remained in place, and the inference was possible when the object behind it exited from its side (Potential deduction phase; movies S3 and S4). Thanks to this modification, the sequences of objects visible to the infant in the inconsistent/consistent outcome phases was reversed with respect to Experiment 1, thus allowing us to control for several non-logical explanation of the results of Experiments 1-2.

Participants

Twenty-four healthy full-term 19-month-old infants (*M*= 19m 8d, range 18m 15d - 20m 2d) were retained for the analyses (12 girls). An additional 6 infants were tested but not retained due to fussiness (4), caretakers' interaction (1), experimenter error (1). Participants were recruited as in the previous experiments.

Materials

The materials of Experiment 3 were identical to Experiment 2 except for the changes specified below. First, the number (7) and the content of familiarization movies were adapted to the nature of the events in the test movies.[3] Second, in the test movies (24) the occluder never lowered to reveal the object behind it. Instead, the object exited from its right edge, stopped between the occluder and the cup, executed its characteristic movement, and after 2s returned behind the occluder. After 1s, another object exited from the occluder. Half of the times, it was the same object (*Consistent Outcome)*, whereas in the other half it was the second object of the couple (*Inconsistent Outcome*). The total

---

[3] In familiarization movie 1 and 2, one object (a blue puppet in movie 1 and a blue flower in movie 2) was on scene, scooped by the cup, moved to the left region of the scene and eventually revealed. Afterwards, the occluder covered the left empty region of the screen and then was removed. In familiarization movie 3 and 4, one object (the blue puppet in movie 3 and the blue flower in movie 4) was on scene, then covered by the occluder. Afterwards the cup scooped the object behind the occluder, it moved it to a different location and eventually revealed it, while the occluder stayed in place. In familiarization movie 5 and 6, two objects (the blue puppet and the blue flower) were on scene, then covered by the occluder. Afterwards, one of the two objects (the blue puppet in movie 5 and the blue flower in movie 6) exited from the right edge of the occluder. In the familiarization movie 7, the six test objects lied side by side and moved one after the other to attract the viewer's attention.

length of each test movie was 17.93 s. The overall occlusion length was 6.82 s. The test movies resulted from the combination of outcome type (Consistent/Inconsistent), pair of objects (Doll-Umbrella/Flower-Dinosaur/Star-Triangle; fig. S1, A, B and C), object scooped by the cup (Star/Triangle/Flower/Dinosaur/Doll/Umbrella) and initial distance of the scooped object from the cup (Near/Far).

Procedure

The procedure of Experiment 3 was identical to the procedure of Experiment 2 except for the following changes. First, there were 7 familiarization trials as explained above. Then, the administration of the test conditions within and between subjects was identical to Experiment 1 except for the fact that the object hidden inside the cup was never revealed. The order of presentation of the final objects reappeared was counterbalanced between participants.

Results

Rejection criteria were identical to those of the previous experiments. Nine percent of the inconsistent and 6% of the consistent trials were excluded by the median filter. An ANOVA as in Experiment 2 revealed that, despite the change in movie structure, nineteen-month-olds looked longer at the inconsistent than at the consistent outcome ($M_{Consistent}$ = 4.9 s, $M_{Inconsistent}$ = 6.2 s; $F(1, 23) = 8.5$, $P = 0.008$). A non-parametric analysis confirmed this result (Wilcoxon matched-pairs Signed-rank, $z = -2.54$, $P = 0.011$, two-tailed).

Experiment 4

Experiment 4 adapts the stimuli designed for Experiment 3 to 12-month-old infants. Apart from changes in the material, the questions asked and the conceptual structure of the stimuli do not change.

Participants

Twenty-four healthy full-term 12-month-old infants (M= 12m 2d, range 11m 22d - 12m 13d) were retained for the analyses (13 girls). An additional twenty-three infants were tested but not retained due either to fussiness (10), caretakers' interaction (5), experimenter error (4) or insufficient valid samples (4).

Materials

The materials of Experiment 4 were identical to those of Experiment 3 except for the changes reported below. There were 8 familiarization movies[4] and 16 test movies, identical in structure to those of Experiment 3. As with our previous experiment with 12-month-olds, pilot studies suggested to use only two pairs of objects. We retained the Doll/ Umbrella couple from Experiments 1-3 and added a novel, Snake/Ball, pair (fig. S1, A and E). The duration of the test movies was 19.43s (see movie S3 and S4).

Procedure

The procedure of Experiment 4 was identical to Experiment 2 except for the use of 8 familiarization trials, designed to introduce participants to the subcomponents of the scenes without ever showing the test scene (Table S1). They were presented in the following order: 1-2 (counterbalanced), 3-6 (random), 7-8 (counterbalanced). In the test phase, participants saw four movies. Order of presentation (CIIC/ICCI), initial distance of the scooped object from the cup (NFFN/FNNF), and object pair order (1212/2121) were counterbalanced across participants.

Results

Rejection criteria were identical to those of previous experiments. Nine percent of the inconsistent and 20% of the consistent trials were excluded by the median filter. An ANOVA identical to Experiment 2 revealed that, despite the change in movie structure, twelve-month-olds looked longer at the inconsistent than at the consistent outcome ($M_{Consistent}$ = 4.2 s, $M_{Inconsistent}$ = 6.1 s; $F(1, 23) = 11$, $P = 0.003$). A non-parametric analysis confirmed this result (Wilcoxon matched-pairs Signed-rank, $z = -2.86$, $P = 0.004$, two-tailed).

Experiment 5

Experiments 1-4 presented infants with situations in which the identity of a partially hidden object could be determined on the basis of a disjunctive syllogism. There, the measure of success was given by looking times at outcomes either compatible or

---

[4] In familiarization movie 1 and 2, one object (a red puppet in movie 1 and a red star in movie 2) was on the scene, then covered by the occluder and then revealed once the occluder was removed. Afterward, the cup scooped the object, it moved it to a different location and eventually revealed it again. In familiarization movies 3, 4, 5 and 6, one object (in each movie, a distinct one of the objects then used in the test) was on the scene, then covered by the occluder. Afterward, the cup scooped the object behind the occluder, it moved it to a different location and eventually revealed it. In familiarization movie 7 and 8, two objects (the red puppet and the red star) were on scene, then covered by the occluder. Afterward, the cup scooped one of the objects behind the occluder, and it moved it to a different location. Eventually, one object exited from behind the occluder (the red puppet in movie 7 and the red triangle in movie 8), while the other object remained hidden inside the cup.

incompatible with the conclusion of an inference. Experiments 5-6 compare the same reactions to the same outcomes when no inference is needed. This can be obtained by creating scenes in which the Potential Deduction phase remains unchanged, but participants need not draw any deduction to determine the content of the cup because they had direct perception of which object entered it. We realized these scenes by inverting the order in which the scooping of one object and the rise of the occluder occurred in the scenes of Experiments 3-4, so that the cup scooped one object in full view, before the occluder was raised.

While it is expected that infants look longer at inconsistent outcomes in these conditions, given what we know about how infants at these ages react at physical inconsistencies, the crucial measure of interest for our analysis will be what happens *before* the outcomes. That is, the main purpose of Experiments 5 and 6 will be to compare the temporal course of infants' oculomotor responses in the Potential Deduction phase when an inference is required (as in Experiments 1-4) and when it is not (as in Experiments 5-6). The analysis of looking times at the scene's outcomes, in a VOE paradigm, is needed in order to validate our procedure and check that infants do react at the outcomes of the scenes as expected.

Participants

Twenty-four healthy full-term 19-month-old infants (*M*= 19m 21d, range 19m 6d - 20m 6d) were retained for the analyses (15 girls). One additional infant was tested but not retained due to insufficient valid samples (1).

Materials and Procedure

The materials of Experiment 5 were identical to those of Experiment 3 except for the test movies. In them, the cup scooped one object in full view and then occluder was raised. Thus, the infants always knew the content of the cup, without the need to make any inference. This modification added 1s to the duration of the movies before the Potential Deduction phase, but starting from it, the movies were identical to those of Experiment 3. The procedure was identical to that of Experiment 3.

Results

Rejection criteria were identical to those of previous experiments. 3% of the inconsistent and 17% of the consistent trials were excluded by the median filter. A preliminary two factor ANOVA with object pair as a within-participant factor detected a main effect of pair. Therefore, we maintained this factor in the main analysis. The ANOVA with outcome type and object pair as a within-participant factor confirmed the main effect of object pair ($M_{\text{Star-Triangle}}$ = 5.8 s, $M_{\text{Dinosaur-Flower}}$ = 5.6 s, $M_{\text{Doll-Umbrella}}$ = 7.5 s; $F(2, 75) = 3.8$, $P = 0.027$). More importantly, it revealed a main effect of outcome type ($M_{Consistent}$ = 3.8 s, $M_{Inconsistent}$ = 8.2 s; $F(1, 75) = 31.44$, $P = 0.0001$). Crucially, no interaction occurred, showing that, as in Experiment 2, 19-month-olds looked longer at the inconsistent than at the consistent outcome regardless of the objects on display. A non-parametric analysis confirmed this result (Wilcoxon matched-pairs Signed-rank, z = -3.71, P = 0.0002, two-tailed).

Experiment 6

Experiment 6 adapts the stimuli designed for Experiment 5 to 12-month-old infants (Movie S5 and S6). Apart from changes in the material, the questions asked and the conceptual structure of the stimuli do not change.

Participants

Twenty-four healthy full-term 12-month-old infants (M = 12m 01d, range 11m 17d - 12m 18d) were retained for the analyses (11 girls). Sixteen additional infants were tested but not retained due to fussiness (2), crying (1), caretakers' interaction (1), experimenter error (5) or insufficient valid samples (7). Participants were recruited at the UPF laboratory (30) and at the BabyLab at Central European University, Budapest (10), where they were tested with an identical setup.

Materials and Procedure

The materials of Experiment 6 were identical to those of Experiment 4 except for the fact that, as in Experiment 5, the cup scooped one object before the occluder was raised, in full view of the infants. This modification added 1 s to the duration of the movies before the Potential deduction phase. Starting from it, the movies proceeded identically (Movie S8). The procedure of Experiment 6 was identical to that of Experiment 4.

Results

Rejection criteria were identical to those of previous experiments. Twelve percent of the inconsistent and 5% of the consistent trials were excluded by the median filter. Because participants were tested in Barcelona (N = 18) and in Budapest (N = 6), we added this factor into our analyses. A mixed ANOVA with outcome type as within-participant factor, laboratory as between-participants factor, participant as random variable and mean looking time (averaged per condition) as the dependent variable revealed that, 12-month-olds looked longer at the inconsistent than at the consistent outcome ($M_{Consistent}$ = 4.9 s, $M_{Inconsistent}$ = 6.2 s; $F(1, 23) = 4.9$, $P = 0.037$), but no other effects. A non-parametric analysis revealed a strong trend consistent with this result (Wilcoxon matched-pairs Signed-rank, $z = -1.8$, $P < 0.072$, two-tailed).

Analysis of Infants' Oculomotor Responses

Experiments 1-6 show that infants look longer at outcomes inconsistent with an inference than at outcomes consistent with it regardless of the physical realization of the scenes (which varies considerably between Experiments 1-2 and Experiments 3-6). Experiments 5-6 show that this looking behavior is comparable to situations in which infants have access to the full information about the identity of objects, without the need of drawing an inference, suggesting that a mental deduction has the same role in fixating infants' beliefs as direct experience. However, the main aim of our research is to explore behavioral correlates of inference drawing *online*. To this purpose, eye tracking data during the Potential Deduction phase, as well as measures that relate infants' behavior during this phase with their response to outcomes inconsistent with a potential deduction,

are crucial. Below, we analyze pupil change, eye position onscreen and proportion of of trials in which infants' eyes shifted from the visible object onstage to the cup in perceptually identical scenes that either invited or did not require a logical deduction (respectively, Inference condition, Experiment 3 and 4 and No-inference condition, Experiment 5 and 6).

The meaning of these variables can be understood as follows. Pupil dilation is the most physiologically meaningful variable. Dilation is dependent upon several factors, partly endogenous and partly exogenous. The fact that we measure it when participants are looking at the same stimulus should minimize changes unrelated with endogenous factors. Cognitively, pupil dilation has been found to be a marker for cognitive effort, memory load and attentional focus (*29*). Shifts from the visible object to the cup are an indication that infants deploy a particular a scanning strategy during inference making. It is particularly worth considering that under every account in a stimulus with a colorful rich moving object and a uniform still cup, the object should attract infants' attention most. Thus, shifts from the superficially most interesting to the less interesting element of a scene may indicate that in that particular moment the less interesting element assumes a particularly important role in the infants' mental processes.

While potentially revealing of infants' strategies, in the short period of the Potential Deduction phase only very few shifts can occur, and in our data generally at most one. For this reason, this variable is not well suited to study the temporal dynamics of infants' oculomotor behavior. Mean *x* gaze position onscreen can do that. This is why we included it in our analysis. However, the reader has to bear in mind that such variable has no direct physiological meaning. Mean *x* describes the variation of eye position onscreen as if it eyes were smoothly transitioning across different elements of the screen, but this is certainly not the case when the objects are still onstage. The apparent smooth transition is an artifact of averaging across different participants who move their eyes in saccades at different times. What mean *x* gives, though, is a good proxy for the temporal course of infants' strategies, as a more extreme mean *x* towards the cup during the Potential Deduction phase indicates when infants tend to focus on the object hidden in the cup. For this reason we analyzed that. Finally, obviously mean *x* and object-to-cup shifts are strictly dependent on the details of the realization of the scenes infants see, whereas (barring luminance differences between stimuli) pupil dilation does not.

Participants and materials

Infants' eye gaze data from Experiments 3-6 ($N = 96$) were analyzed. Trials in which less than 70% of the gaze samples were collected were excluded from analysis. With these criteria 4 participants were excluded (1 in Experiment 5 and 3 in Experiment 6) because no valid trial was retained. All temporal analyses were performed with DataDesk 7 Statistical Analysis software (https://datadescription.com) and Matlab R2015b (http://www.mathworks.com/).

Results

In our analysis of oculomotor responses, we focused on the Potential Deduction phase. This started when the object was visible while exiting the occluder, spanning

through the subsequences as in Fig. 3A, and ended at the last frame before the final object exited the occluder.

The Temporal Dynamics of the Potential deduction phase: pupil change and mean x gaze position.

We analyzed the temporal dynamics of the mean *x* gaze coordinates (hereafter mean *x*) and the pupil changes in the same temporal region. Pupil dilation were calculated as follows. First, we identified the first frame at which the mean *x* gaze did not differ statistically between the two conditions (Inference and No-Inference, unpaired *t-test*). The rationale for this choice is that starting from that frame we could be sure that the eyes were directed at the same area of the screen, and hence that the pupil was exposed to identical light sources in both conditions. Thus, any potential luminosity difference between conditions could not affect the dilation of the pupil at the relevant moment of the analysis. This choice was facilitated by the fact that the Potential Deduction phase started when an object exited the occluder. The sheer movement of the object onscreen led most participants in most trials to direct their gazes towards the exiting object, marking a quite clear moment in which their *x* gaze position would overlap, on average. Then, for each trial and for each participant, a baseline pupil diameter was computed by averaging the data samples in the first 300 ms after that frame. Such interval should be sufficient for the physiological response of the pupil to adapt to the lighting condition of each movie at the moment of interest. Second, for each successive data sample during the Potential Deduction phase, we computed the difference between raw pupil diameter and the baseline pupil diameter, trial by trial and participant by participant. Thus, pupil dilation changes reflected actual differences at each individual trial for each individual infant. Finally, for the pupil dilation changes as well as the mean *x*, we ran a cluster mass test coupled with a randomization procedure (*24, 25*).

The details of our procedure were as follows. We computed unpaired t-tests at each time bin for the mean *x* gaze coordinates and the pupil dilation changes, separately for 12- and 19- month-olds. For pupil dilation changes, the tests were one-tailed because, based on our hypothesis and the literature, we expect an inference to have a cognitive cost resulting in higher pupil dilation, as for other cognitive activities implying higher effort. For mean *x*, the tests were two-tailed we have no hypothesis as to where infants look in the particular moments of the Potential Deduction phase we analyzed. Temporal cluster statistics were defined as the sum of the *t* values thresholded at p=0.05 on consecutive time bins. To evaluate the significance of the test, we recomputed the same analysis on 1000 sets of random permutations. The permutations were computed by randomly assigning the participants to the condition labels (Inference/No Inference), taking care to maintain the N of the two conditions identical to the N available for the original data.

Twelve-month-old infants results. In the short period during the Potential Deduction phase, 12-month-olds' pupils dilated more in the Inference than in the No-inference condition. This occurred for a temporal window of about 370 ms (Fig. 2E).

Remarkably, this effect occurred when participants were looking at the same area of the screen, suggesting that it could not be due to differential luminosity at the fixation point.

Close to the end of the Potential Deduction phase (2.3 s after its beginning), infants in the Inference condition also tended to look more towards the cup, for a cumulative time slices totaling about 370 ms, as revealed by the mean *x* analysis. This occurred when only the occluder and the cup containing the scooped object were on stage (Fig. 3A*v*), suggesting that the displacement towards the cup was a sign of a scanning strategy deployed by infants during the inference. We will further analyze the relevance of this result below.

Nineteen-month-old infants results. While the temporal details of the Potential Deduction phase were slightly different between 12 and 19-month-olds (Fig. 3A), the results were pointing, even more markedly, towards the same conclusion: infants' pupils dilated more in the Inference than in the No-inference condition. This occurred in consecutive time slices totaling about 1.2 s (Fig. 3C). Again, this occurred even when participants in the two conditions were looking at the same screen areas, ruling out the possibility that variations in pupil response depended on differential luminosity at the fixation point. As in 12-month-olds, infants' eyes tended to look more towards the cup around the end of the Potential deduction phase (1.6 s after its beginning), for about 630 ms.

It can be noted that, overall, in 19-month-olds the pupil tends to get smaller across the Potential Deduction phase. The same decrease cannot be observed in 12-month-olds. To understand the reason for this difference, consider again the critical sequence of actions of the Potential Deduction phase (Fig. 3A). First the object exits the occluder (*iv-a*). Then it stays in place and wiggles to attract infants' attention (*iv-b*). Then it returns behind the occluder (*v*). Pilot studies suggested to lengthen this phase for 19-month-olds, because they disengage faster from the stimulus than 12 month-old infants. Thus, object animation in *iv-b* lasts longer for 19-month-olds than for 12-month-olds. And because in this sub-phase, beside object movement nothing cognitively noticeable occurs, the pupil in 19-month-old infants may have decreased faster. Indeed, also in 12 month-old infants this phenomenon can be observed during the same sub-phase; only, this phase is shorter and the effect is less visible. In short, the differences in pupil size reduction across the Potential Deduction phase may be attributed to differences in the lengths of the sub-components of that phase, introduced in order to adapt the stimuli to the overall reactions of infants of the relevant age. Importantly, our conclusions are based on differences in the dynamics of pupil dilation between conditions, and these are preserved even if despite differences in the physical realization of the scenes.

Infants' scanning strategies: obiect-to-cup shifts during the Potential Deduction phase

To study infants' scanning strategies during the Potential Deduction phase, we checked the number of trials in which infants switched their gaze from the visible object to the cup during the Potential Deduction phase. We defined two rectangular areas of interest in the stimuli window, one centered around the position of the object exiting the occluder after it stops by the cup, and one centered around the visible top-part of the

object hidden inside the cup (Fig. 2, *iv*). A shift was defined as a passage from the first to the second area, identified with two successive gaze positions in the two distinct areas, during the time window in the Potential Deduction phases *iv-b, v-a, v-b* (Fig. 3 A).

We ran a two-way ANOVA with proportion of trials as the dependent variable, Age Group (12/19) and Condition (Inference/No-Inference) as an independent variables. As reported in the main text, the analysis revealed an effect of Condition, no effect of Age and no interaction. Infants shifted from the object to the cup in more trials when the Potential Deduction phase required an inference than when it did not ($M_{Inference}$ = 71%, $M_{No\_Inference}$ = 50%; $F(1, 88) = 10.4$, $P = 0.002$; fig. S2). This result shows that in the brief period of the Potential Deduction phase infants clearly deploy different scanning strategies depending on whether the content of the cup is known or not.

The predictive value of oculomotor responses: the relation between oculomotor behavior during the Potential Deduction phase and looking time in the Outcome phase.

An important question is the nature of the differential effects we found in the Inference and No-Inference condition across the three oculomotor variables we examined. We can exclude that these differences are due to the fact that infants are looking at different scenes at the moment of the Potential Inference phase, which is identical in both condition. However, we cannot exclude that the different time course of oculomotor patterns may be the result of processing event structures prior to the Potential Deduction phase, rather than logical inferences. In this case, the dilation patterns serve as markers of event binding as a result of the time structure in the stimuli. For example, the action of occluding objects prompts changes in attention and memory (and concomitant increases in pupil dilation) as does the presentation of new information after occlusion. Under this interpretation, differences between Inference and No-Inference conditions are artifacts of the memory of past events, or of how they affect event binding at the moment of the Potential Deduction phase.

To explore this possibility, we checked whether the oculomotor patterns at the moment of the Potential Deduction are predictive of the looking time behavior that infants manifest at the end of the scenes.

To do this, first we computed a simple measure of success at detecting inconsistent outcomes in the Outcome phase. We operationalized it as the differences between looking time at the inconsistent minus looking time at the consistent outcomes, computed per each infant. We then extracted a single measure of pupil change and mean *x* per each infant, by averaging the values of these variables in the bins where the maximum differences between conditions occurred, as revealed by the respective cluster mass tests (Fig. 3). We then transformed the variables into z-scores, computed separately per each experimental group, and used such values as the basis for our analysis.

We regressed the normalized measure of success in the Outcome phase against the normalized oculomotor measures in the Potential Deduction phase, separately for infants in the Inference and the No-Inference conditions because our prediction is that magnitude of the oculomotor response at the moment of the inference should be predictive of surprise when an inference can be computed, but not otherwise. We studied a least square

sets of models. We first identified potential data points unduly influencing the regression. For infants in the Inference condition, a Potential-Residual plot after Hadi's influence measure signaled four potential outliers. We added them individually as separate factors in the regression. Two of these points highly influenced the model, and were thus excluded from further analysis. The same procedure identified two outliers in the No-Inference condition, which were also excluded from analysis. With the remaining participants, we first explored the relations between oculomotor measures to check for collinearity. Average pupil change was poorly correlated with either mean *x* or proportion of shifts (respectively, $r = -0.2$ and $r = 0.18$; Pearson Product-Moment Correlation). However (understandably) mean *x* and proportion of shifts were highly correlated ($r = 0.68$). Therefore, we ran two final models: in one we kept mean pupil change and mean *x* as predictors, and in the second one we kept mean pupil change and proportion of shifts as predictors.

For participants in the Inference condition, the model that accounted for most variance was the one regressing success in the Outcome phase against mean pupil change and proportion of shifts in the Potential Deduction phase. These factors accounted for the 19% of the variance (R squared adjusted = 18.9%, $F(2, 43) = 6.4$) and both were significant positive predictors ($t_{\text{Mean Pupil Dilation}} = 2.0$, $P = 0.045$; $t_{\text{Proportion of Shifts}} = 3.2$, $P = 0.002$). For participants in the No-Inference condition, no model revealed any relation between any oculomotor measure and success in the Outcome phase. We report the raw, the z-scores and the partial correlation plots in fig. S3.

We compared the results of the least square model with a robust regression model based on the Huber regression, not requiring outlier selection. We fit two models to our data, one for inference and one for no inference condition, using the same predictors used for the least square regression analysis. In the Inference condition, proportion of shift was a significant coefficient ($c = 0.26$, $P = 0.05$). Pupil tended to influence the regression, but to a lesser degree, and did not reach significance ($c = 0.17$, $P = 0.14$) Again, in the No-inference condition no predictive variable showed any relationship with success in the Outcome phase.

In sum, these regression techniques showed that proportion of shifts during the Potential Deduction phase predict success at identifying inconsistent outcomes in the Outcome phase, as did, to a lesser degree, pupil change. However, these relation occurred only for those infants who were in the Inference condition. By contrast, for infants in the No-inference condition, oculomotor data have no explanatory power on surprise at the inconsistent outcomes. Thus oculomotor responses in the Potential Deduction phase are indeed related to looking time at the outcomes, and are a function of the logical status of the otherwise identical scenes perceived by infants. The fact that this relation is predictive only when an inference is involved suggest that these measures are neither simple markers of event binding, nor simple reflexes of memory traces for the past unfolding of the scenes prior to the Potential Deduction phase. Whatever theory best explains what infants think in the Potential Deduction phase, our data license the conclusion that oculomotor responses during that phase are tied to the computations that generate infants' expectations for the final outcomes. These computations are modulated

by the the need to make an inference to identify the content of the object hidden in the cup.

In conclusion, the analyses of oculomotor responses showed that, remarkably, at both ages three markers of inference making clearly appeared. We found higher pupil dilation, as signaled by the analysis of the temporal course of the Potential Deduction phase, which is a potential sign of processing cost. We also observed a reorientation of focal attention toward the cup, as signaled by the temporal analysis of mean *x* and by the analysis of the proportion of trials with visible-object-to-cup shifts. Such oculomotor markers confirm that the longer looking time at the outcomes of scenes inconsistent with a deductive inference, as found at the outcomes in Experiments 1-4, were likely due to the fact that infants in the Inference condition made an hypothesis about the content of the cup as early as the information to draw the deduction licensing it was available, right during the Potential Deduction phase (see fig. S4 for a compact representation of the oculomotor responses during the full Potential Deduction phase).

Experiment 7

The results of Experiments 1-6 strongly suggest that, regardless of the physical realization of the scenes and of age, infants draw elementary logical inferences online and use them to predict future outcomes. This inferential activity was accompanied by remarkably stable behavioral signs, affecting cognitive load during inference making and scanning patterns of the scenes. These results raise the issue of whether the years of experience with language and the enormously increased knowledge of the world that adults possess compared to young infants could change how inferences are processed and what behavioral signs reveal inference making when adults inspect nonverbal scenes with the same logical valence as those we presented to infants. To this purpose, we tested whether the same behavioral signs of inference making could be detected when adults passively look at movies similar to the ones used in Experiments 1-6, when they have no explicit logical task. A second interest of this research is that the study of nonverbal logical reasoning in adults is a virtually untouched domain. Comparing how infants and adult react to structurally similar visual material may advance our knowledge of nonverbal reasoning processes in humans.

Participants

Thirty participants were retained for analysis (*M*= 23 yr 9d; range=18:30; 22 females). Another four participants were excluded due to total loss of valid data gaze points. Participants were recruited via the participants' database of the Universitat Pompeu Fabra. They were paid € 10 for their participation. All participants had normal vision and had no neurological or psychiatric history. After being instructed about the procedure, they gave their informed consent.

Materials and Methods

Stimuli were animation movies similar to the ones used in Experiments 1-6. They were generated with the software Apple Keynote as QuickTime movies, adjusted for timing, compiled at 60 fps and compressed with Apple Intermediate Codec format.

Unlike the movies in Experiments 1-6, no sound was added to the videos. In half of the movies, during the Potential Deduction phase the occluder was lowered and the content of the cup was revealed at the end of the scenes (as in Experiments 1 and 2). In the other half, the object exited from behind the occluder and the content of the cup was never revealed (as in Experiments 3-6). These movies were crossed so as to yield three types of scenes: *Inference, No-inference* (of type 1 and 2, as explained below)*,* and *Interruption scenes*.

Inference scenes (*N*=32) were exactly as the test movies in Experiments 1-4. In them, an inference was needed in order to determine the content of the cup (Movie S7). In *No-inference scenes of type 1* (*N*=32), as in the test movies in Experiments 5 and 6 (Movie S8), the cup scooped the object before the occluder raised, so that its content was always known and no inference was needed to determine the identity of the object inside it. *No-inference scenes of type 2* (*N*=32) were identical to test inference scenes, but the objects were paired in such a manner that their upper part was different. This was obtained by yoking the couples with identical upper parts; for example, the snake was paired with the umbrella (Movie S9). Once the cup scooped the object from behind the occluder, its visible part provided enough evidence to disambiguate between alternatives. Thus, even in this case, no logical inference was needed to determine the content of the cup. The role of No-inference scenes of type 2 was twofold. First, they added variability in the stimuli, making it more difficult for participants to focus on parts of the scenes in order to determine when they had to make an inference. Second, most importantly, they allowed us to control for differences in the unfolding of the scenes prior to the Potential Deduction phase. Type 1 No-inference scenes are matched for object properties to their Inference counterparts, but not for the action sequences prior to the Potential Deduction. Indeed, (Type 1) No-inference scenes are obtained by Inference scenes by inverting the initial sequence of actions, with the scooping occurring before the occlusion. By contrast, Type 2 No-inference scenes are not matched for object properties with their Inference counterparts, but the action sequence before the Potential Deduction phase is exactly identical to that of Inference scenes (e.g., Fig 2A-B). In all cases, Inference scenes were matched with corresponding No-inference scenes so that they shared identical Deduction phase. Thus, for every Inference scene there were two No-inference scenes (of Type 1 and 2) associated to it, identical in the crucial part to be analyzed. In all cases, in order to reproduce the conditions experienced by infants as close as possible, half of the scenes ended with a consistent outcome and half with an inconsistent outcome. The scenes ranged in length from 16.680 to 20.480 ms. Again, such differences were introduced in order to avoid that participants would focus only on specific parts of the scenes, without considering the entire sequence of events.

Interruption scenes (*N*=48) were identical to Inference and No-inference scenes, but in particular points they froze and did not continue. They too were added in order to ensure that participants could not precisely predict which scenes required an inference and which ones did not. Interruption scenes were associated to a memory task, as explained below.

After the presentation of a scene, participants were asked to punch in a response by selecting a picture on screen with a mouse click. The type of response was different for Inference/No-inference scenes and for Interruption scenes (See *Procedure)*. All movies were presented centered on the 24" monitor incorporated into a Tobii XL60 eye tracker, which also recorded eye data at 60 Hz. They were scaled to a 34.27 x 24.5 cm size. The experiment was controlled by the software Psyscope X. An oval table was placed between the participants and the monitor, with a wireless mouse (Apple Magic Mouse 2) that could register participants' responses when the procedure required it. The experimental room remained illuminated during the entire session, at an average of 5.5 lx.

Procedure

Participants sat on a fixed chair, approximately 70 cm away from the monitor. At the beginning of the experiment they received verbal and visual instructions. They were told that they would see a set of movies, some of which contained errors. They were informed that the movies had been mixed up among good and bad movies, and that they had to help in separating them by indicating which ones were correct and which ones were incorrect. They were also instructed that sometimes the videos would stop and they would be given a test about the cup content. There was no mention that they would perform a task of logical reasoning. After receiving the instructions, participants underwent a nine-point calibration procedure programmed in PsyScope X. After calibration, the experiment proper started.

Participants saw 96 Inference and No-Inference movies, and 48 Interruption movies in a within-participant design. Movies were presented pseudo-randomly, with the following constraints: (I) There were at most five consecutive Inference/No-Inference or Interruption movies; (II) There were at most three consecutive scenes of Inference or No-inference kind; (III) There were at most three consecutive Cup content revealed or Cup content not revealed movies; and (IV) There were at most three consecutive Consistent or Inconsistent outcome movies. At the end of the movies, participants had to judge whether the scene was correct by clicking on a multiple-choice panel presented on screen. After the Inference/No inference movies, two pictures of a green checkmark or a red cross were presented, so that participants could indicate whether they thought that the movie was correct or incorrect. After the presentation of an Interruption movie, four pictures depicting the content of the cup were presented. The four alternatives offered were: an empty cup, object A, object B, or a cup with an ambiguous content (either A or B). Participants had to click on the picture that they thought would correspond to the content of the cup. All alternatives were presented pictorially, without any text element. A new scene was displayed once participants had made their choices. A pause of 5 min was enforced at about the 60% of the experiment. The whole experiment took on average 60 min.

Results

Exclusion criteria. For all participants, six test movies were excluded due to experimenter's error. Also participants for whom more than the 30% of the eye tracker data was lost were excluded from analysis (N=4). For the remaining participants, we

excluded from analysis those scenes in which more than 30% of data points in the Potential Deduction phase were invalid or lost (10% of the movies). A gaze was considered to be valid if the validity index recorded by PsyScope X was higher than 3 over a maximum of 4.

Participants' Responses. Responses to the test scenes were graded as correct if participants clicked on the green checkmark for scenes with consistent outcomes or on the red (cross) for scenes with inconsistent outcomes. With these criteria, participants' answers were 97% correct. Answers for interruption scenes were graded as correct if participants identified the content of the cup at the moment of the interruption correctly. For these, participants were 94% correct. These values show that participants followed the scenes correctly and responded to some general sense of their internal coherence, although no explicit logical task was given to them.

Oculomotor Responses. The analysis of main interest is the temporal course of participants' eye gaze and the variation in their pupil dilation during the Potential Deduction phase. As in Experiments 3-6, the onset of this phase corresponds to the moment in which the mean x coordinates converged in both conditions (Inference and No-Inference). We ran both overall analyses of differences between conditions, as well as detailed temporal analyses to better describe the temporal course of a possible inferential process.

For the overall analyses, we ran repeated measures ANOVAs with either Pupil dilation or Mean *x*-coordinates as dependent variables, Condition (Inference vs No-inference) as independent variable, and participants as random factor, in a within-participant design.

For the temporal analyses, we analyzed the interval corresponding to the Potential Deduction phase into the 16.6 ms bins provided by the eye tracker sampling rate. We assessed statistical differences between conditions by using a cluster mass test coupled with a randomization procedure, as in Experiments 3-6. Unlike Experiments 3-6, we ran pairwise *t*-tests because the conditions of interest were run within-participants. As in Experiments 3-6, for the reasons explained therein, tests were one-tailed for pupil dilation and a two-tailed for mean *x* coordinates. The criteria for calculating the pupil baseline were identical to those used for infants, that is, computed at each trial of each participant, so as to adapt baselines to the actual light received by their pupils at each trial. Also the statistical criteria for the time course analyses were identical to those used in infants, except for one aspect of the permutation procedure: because the design of the experiment was within participants, in order to generate the population of permutations, trial labels (Inference/No Inference) were permuted within each participant, keeping the number of trials of each condition identical to the N actually collected for that participant.

Adult participants saw all the kinds of movies and the conditions that the infants saw in Experiments 1-6. Because the temporal and spatial properties of scenes in Experiments 1-2 (e.g., the positions of the objects; or the behavior of the occluder) are different from those in Experiments 3-6, we ran separate analyses for those scenes in which the inference was triggered by the occluder lowering during the Potential Deduction phase (fig. S5) and those in which the occluder remained in place and the inference was

triggered by the object exiting the occluder (fig. S6). For the same reasons, we also separated the analysis of Type-1 scenes, which closely match the scenes that the infants saw, from the analyses of Type-2 scenes, which only the adults saw and which had the main function of controlling for the order of events in the scene sequence.

We begin by presenting the analyses of the scenes where the occluder lowered during the Potential Deduction phase (fig. S5). An ANOVA with pupil dilation as the dependent variable Condition (Inference/No Inference) as independent variable and participants as random factor revealed that participants' pupil dilated more in the Inference than in the No-Inference condition. This was true both for Type-1 scenes [as reported in the main text: $M_{Inference}$ = -0.103 mm, SD = 0.065 , $M_{No\ Inference}$ = *-0.173 mm,* SD = 0.10; $F(1,29) = 12.46$, $P = 0.0014$] and for Type-2 scenes, [$M_{Inference}$ = -0.103 mm, SD = 0.065, $M_{No\_Inference}$ = -0.145 mm, SD = 0.082; $F(1,29) = 7.69$, $P = 0.0096$]. Also, an ANOVA with mean x as the dependent variable and the same independent factors revealed that participants looked more towards the cup in the Inference than in the No-inference condition in Type-1 scenes [$M_{Inference}$ = 902.33 px, SD = 91.47 px; $M_{No\ Inference}$ = 851.20 px, SD = 114.09; $F(1,29) = 6.21$, $P = 0.018$]. No such overall difference appeared in Type-2 scenes [$M_{Inference}$ = 902.33 px, SD = 91.47 px; $M_{No\ Inference}$ = 892.35 px, SD = 93.94 px; $F(1,29) = 0.66$, $P = 0.42$]. Thus, while the two types of scenes may have induced different exploratory behaviors (understandably, given the different event sequences and event properties), both of them apparently taxed participants' cognitive resources equally, demanding more effort when an inference was needed to determine the content of the cup.

In order to better examine the oculomotor responses during the unfolding of the scenes, we studied the participants' looking dynamics during the Potential Deduction phase. We analyzed a time interval of about 2 s after the beginning of the Potential deduction phase (fig. S5,D-F). In Type-1 scenes, the pupil dilated more while participants were looking at Inference than at No Inference scenes for a consecutive period of 1.2 s. Pupil size started differing from the No-inference scenes at about 450 ms after the onset of eye gaze displacement, with an extended effect in time consistent with the physiological response of the pupil. Besides, the adults consecutively looked more towards the cup when the scene required an inference than when no inference was needed for 0.6 s (fig. S5E). In Type-2 scenes, although no strong difference in mean x displacement emerged, the pupils dilated more during the Potential Deduction phase of Inference scenes for a total of about 1.1 s (fig. S5F). The pupils began to differentially dilate at about 500 ms after its beginning, again, consistent with the pupil's physiological response. The fact that increased pupil dilation in the inference condition occurred also when participants were looking at exactly the same areas of the scenes in both condition shows that the dilation could not be explained by luminance changes among conditions, changes which were in any case minimal.

We then analyzed the scenes where the inference was potentially triggered by the object exiting the occluder (Fig. 2,A and B; fig. S6). Even in this case, the adults' pupils dilated more in the Inference than in the No-Inference scenes, both in Type-1 [as

reported in the Main text, $M_{Inference}$ = 0.0387 mm, SD = 0.084; $M_{\text{No\_Inference}}$ = -0.010 mm, SD = 0.091; $F(1,29) = 5.676$, P = 0.024] and Type-2 scenes [$M_{Inference}$ = 0.039 mm, SD = 0.084; $M_{\text{No\_Inference}}$ = -0.026 mm, SD = 0.102; $F(1,29) = 12.40$, P = 0.0014]. Again this was a sign of higher cognitive load in the inference condition, likely due to inference making. Participants looked more towards the cup when an inference was needed in Type-1 scenes [$M_{Inference}$ = 1135.57 px, SD = 47.56; $M_{\text{No\_Inference}}$ = 1113.15 px, SD = 46.44, $F(1,29) = 8.26$, P = 0.0075], but not in Type-2 scenes [$M_{Inference}$ = 1135.57 px, SD = 47.56; $M_{\text{No\_Inference}}$ = 1124.14 px, SD = 41.24; $F(1,29) = 2.06$, P = 0.161], again showing that the eye scanning pattern may depend on the physical configuration of the scene, but not the pupil response, which was always affected by the need to derive a deduction.

The pupils remained larger during the Potential Deduction phase of the Inference than during the same phase of the No-Inference scenes for an interval of 1.2 s in Type-1 (Fig. 3D) and for about 1.7 s in Type-2 scenes (fig. S6D), revealing a prolonged, more intense cognitive load. The temporal analysis revealed extended contiguous temporal regions in which participants tended to look more towards the cup in the inference scenes, both in Type-1 and Type-2, for a continuous stretch of about 1 s in Type 1 scenes (Fig. 3D) and for a total of 530 ms in Type 2 scenes (fig. S6D).

In conclusion, despite the differences in the types of scenes, participants's pupils were always more dilated when an inference was needed in order to determine the content of the cup than when no such inference was needed. This was true regardless of the exact unfolding of the scene, or of the exact way in which the scene provided evidence for an inference, and regardless of the fact that participants had no explicit logical test to perform. In most (but not all) cases, this clear marker of inferential activity was associated to a tendency to direct the gaze towards the cup containing the hidden object more markedly when participants saw scenes potentially containing an inference. The fact that the pupils dilated more in all inference conditions, even when participants did not necessarily direct their eyes towards the cup, is remarkable. It shows the presence of inferential mental processes that cannot be reduced to the particular elements that participant were inspecting, at a particular moment of the unfolding of a scene.

FIGURES AND TABLES

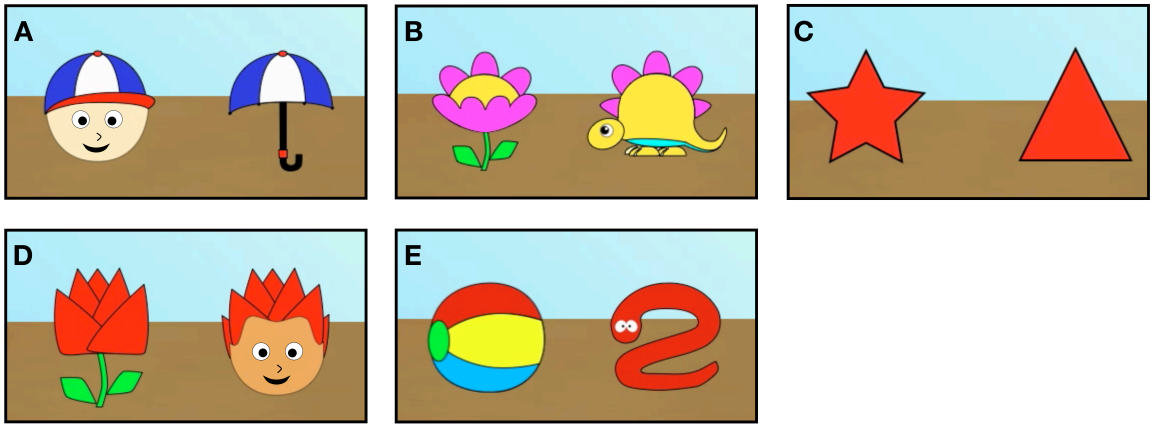

Fig. S1. Objects Pairs used in the Experiments.

The top part of all object pairs is identical, so that the identity of the object hidden inside a cup cannot be determined by visual inspection alone. (**A**) Doll-Umbrella. (**B**) Flower-Dinosaur. (**C**) Star-Triangle. (**D**) Doll-Flower. (**E**) Snake-Ball. For copyright reasons, the images of the dolls in pairs A and D have been modified with respect to the stimuli presented in the relevant experiments, but the changes are irrelevant to the main contrast tested therein.

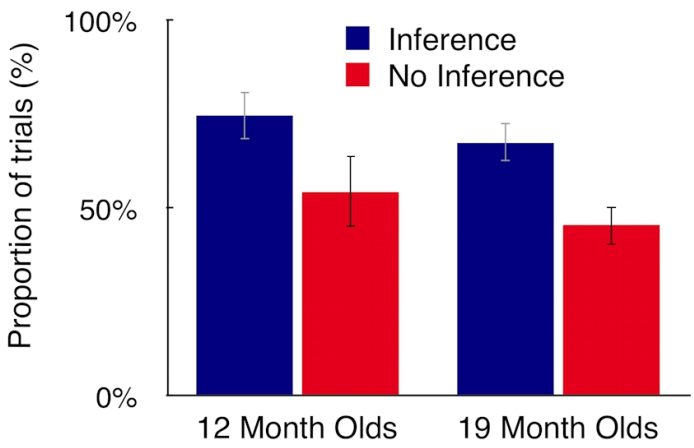

Fig. S2. Proportion of trials with visible-object-to-cup shifts in the Potential Deduction phase (Experiments 3-6).

Bar plots represents the proportion of trials (SE) in which infants shifted from the visible object to the cup during approximately 1.5 s of the Potential Deduction phase. Both 12 and 19-month-old had a higher proportion of trials with shifts when a inference was required to disambiguate the content of the cup than when no inference was required.

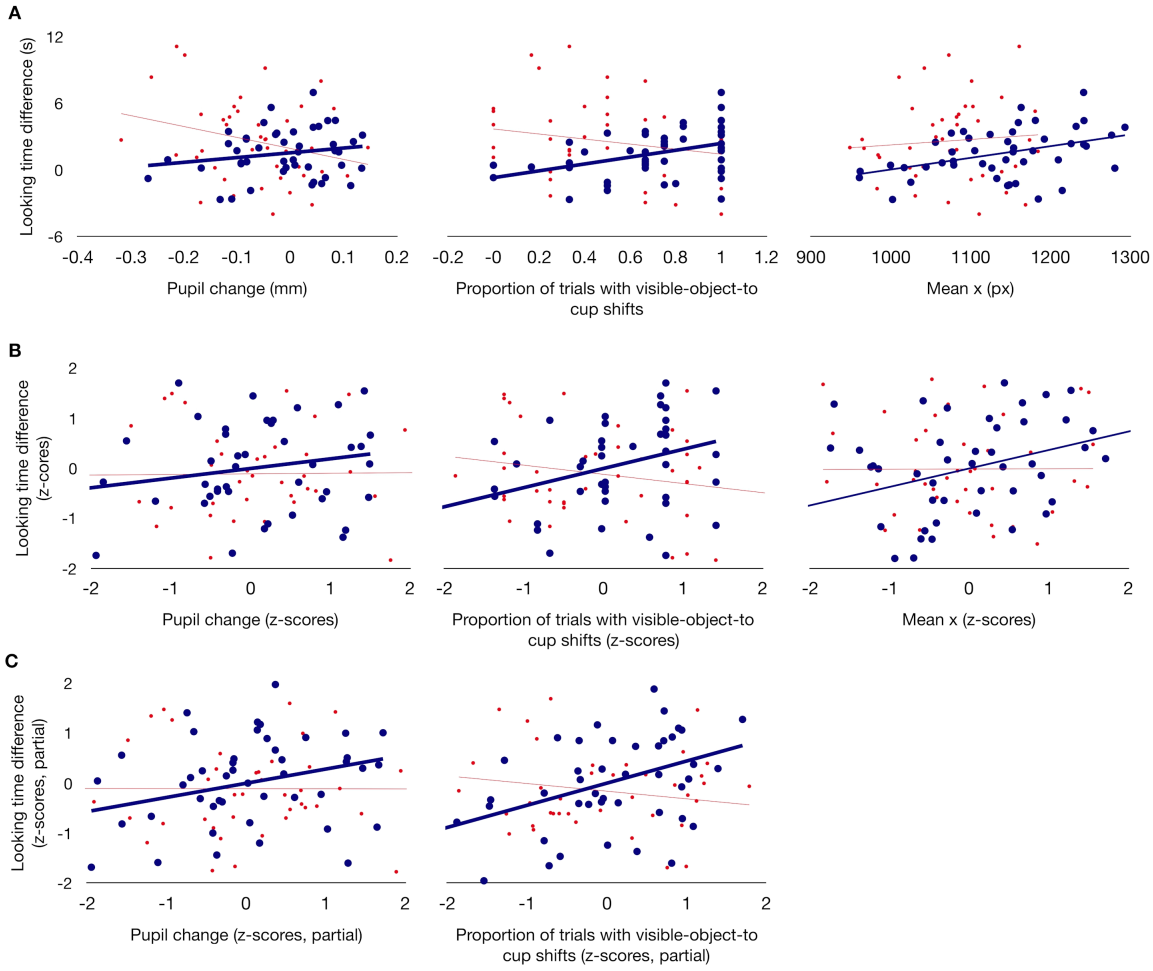

Fig. S3. Infants' oculomotor responses in the Potential Deduction phase are positive predictors of inconsistent outcomes in the Outcome phase, but only when an inference is needed to identify the content of the cup (Experiments 3-6).

(**A**) Scatterplots of the raw data of Pupil change, Proportion of trials with visible object-to-cup shifts, and Mean *x* eye position in the Potential Deduction phase, plotted against a measure of success at identifying inconsistent outcomes (that is, the difference in Looking time between Inconsistent and Consistent trials in the Outcome phase). Thick intercept blue lines mark the intercepts for variables retained in the final regression. Thick blue dots indicate values for infants in the Inference condition. Thin red dots indicate the same values in the No-Inference condition, with an indicative intercept line without statistical significance (**B**) Z-scores of the same data. (**C**) Partial regression plots of the two variables retained in the least square regression which accounted for most variance (R^2 = 22.5%; R^2 adjusted = 18.9%). No model accounted for any variance for the No-Inference participants, whereas Pupil change and Proportion of trials were predictive of

success at correctly reacting to the consistency of the final outcomes for participants in the Inference condition in a least square regression model. Proportion of trials (but not pupil change) was a significant predictor in a robust regression model.

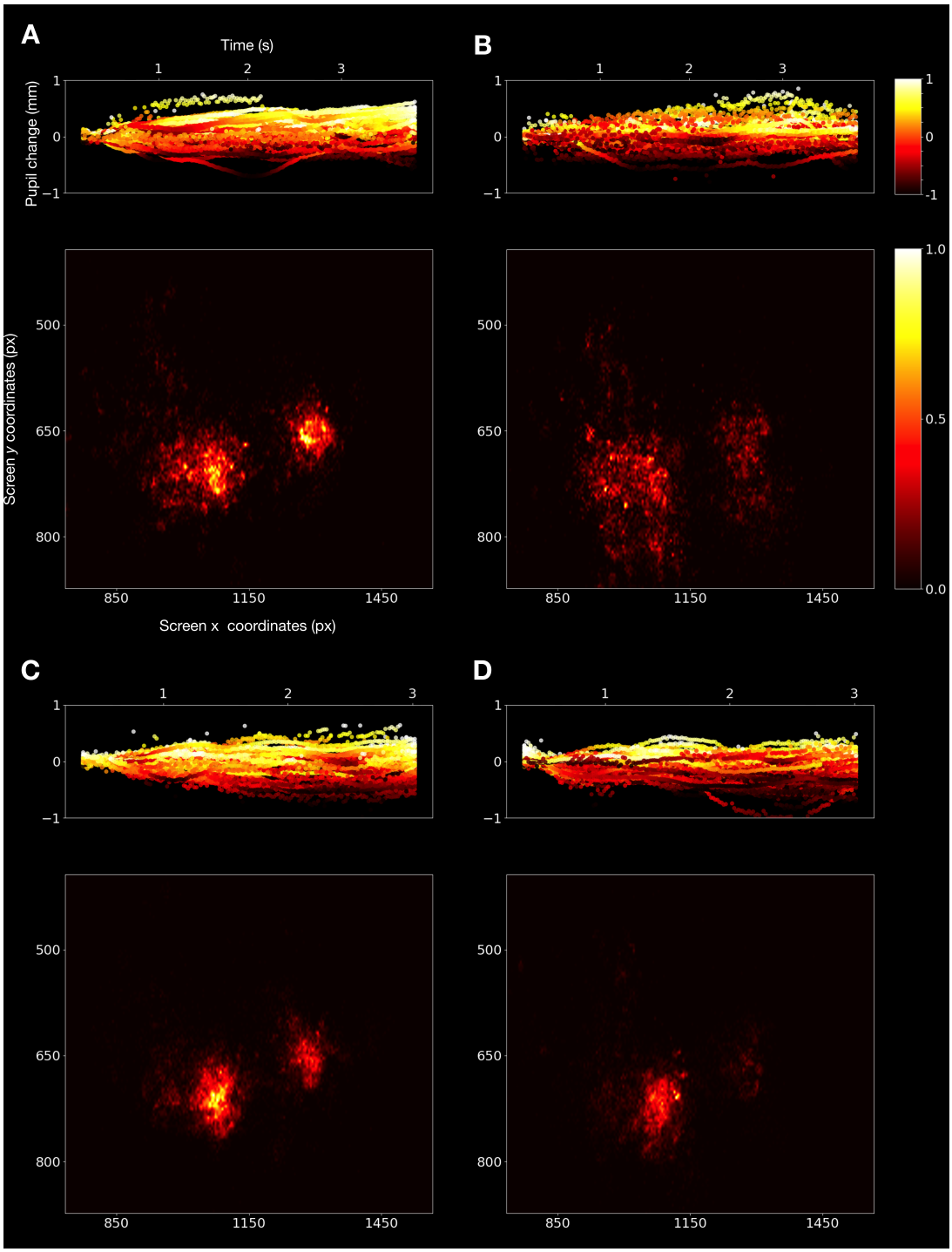

Fig. S4. Time-course of pupil change and density plot of gaze distributions in the Potential Deduction phase.

(**A**) Twelve month-old infants, Inference condition. (**B**) Twelve month-old infants, No-inference condition. (**C**) Nineteen month-old infants, Inference condition. (**D**) Nineteen

month-old infants, No-inference condition. In each panel, the top part represents the time-course of the pupil change of the respective ages. Each line represents an individual trial. The y-axis plots pupil change from baseline (mm). The color represents the amplitude changes in each individual trial, computed per trial. Thus, the lightest color indicates the maximum pupil change during each trial.

The bottom panels represent the density plot distribution of the proportion of gazes across the full Potential Deduction phase. It can be seen that in the two inference conditions (A and C) the maximum gaze density is centered around two separate points on screen, corresponding to the exited object and the cup. By contrast, in the No-inference conditions infants tend to spread their focus on several different parts of the area of interest. It should be noticed that in absolute terms the amount of gazes inside the areas of interest is much higher for the Inference infants than for the No-inference infants, who tend to look at other areas of the screen.

It is also possible to observe that in the pupil timeline plots there is a higher concentration of light colors at the beginning of the phase for 12-month-old infants, and all across the phase for 19 month-old infants, indicating that in a higher number of trials infants in the Inference condition reached their maximum pupil dilation earlier and more consistently.

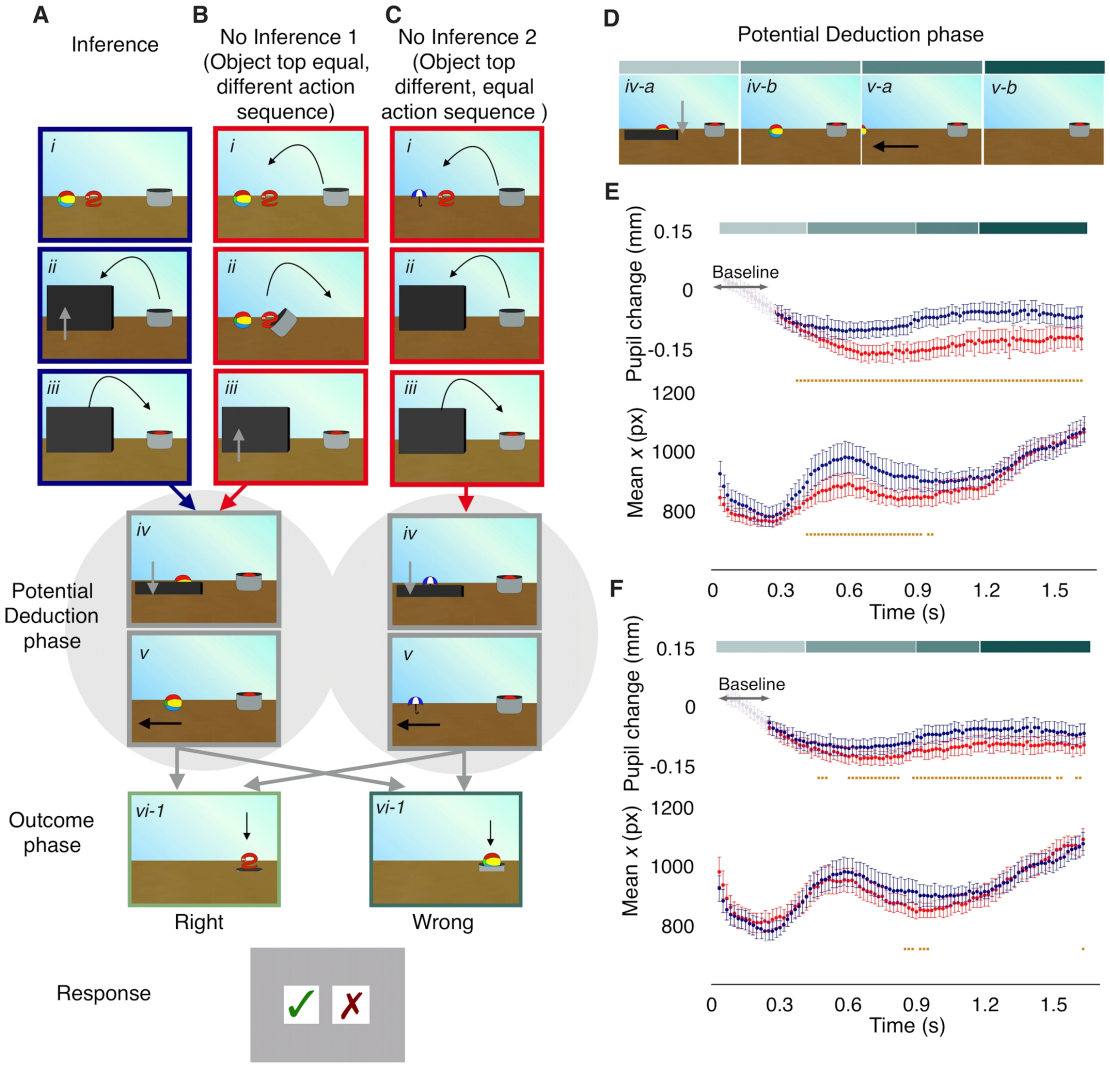

Fig. S5. Scene sequences for Inference and No-Inference scenes of Type 1 and 2 in Experiment 7, and temporal analysis of the Potential Deduction phase (Inference triggered by the occluder lowering).

(**A**) In the Inference Scenes, the object in the cup was unknown, whereas in the No-Inference scenes (**B** and **C**) it was always known, because either the cup scooped it in full view (B) or the objects had different upper parts (C). The Potential Deduction phase of all scenes (**iv-v**) was identical. The only difference was the need of an inference to identify the content of the cup in the Inference, but not in the No-Inference, scenes. In the Outcome phase (**vi**, identical for all scenes) the cup reveals its content, which is either consistent with the logical deduction (iv-1) or inconsistent with it (iv-2). (**D**) Sub-phases of the Potential Deduction phase (**iv-v**), color-coded. (**E** and **F**) Temporal course of pupil

dilation differences from baseline (mm) (top) and Mean *x* gaze positions (px; bottom) for Type 1 scenes (E) and Type 2 scenes (F). Blue and red sample data correspond to the oculomotor behavior recorded during Inference and No inference scenes respectively. Statistical differences were used by computing a cluster mass test with 1000 sets of random permutation generated at each sample of the 60 Hz data flow (one-tailed paired t-tests for pupil and two-tailed t-tests for mean-x position). Yellow dots indicate regions of differences in the two conditions. Adults had higher pupil dilation in Inference scenes in all cases, regardless of where they were looking and of how the No-Inference scenes were realized. Error bars are 95% within-participants confidence intervals.

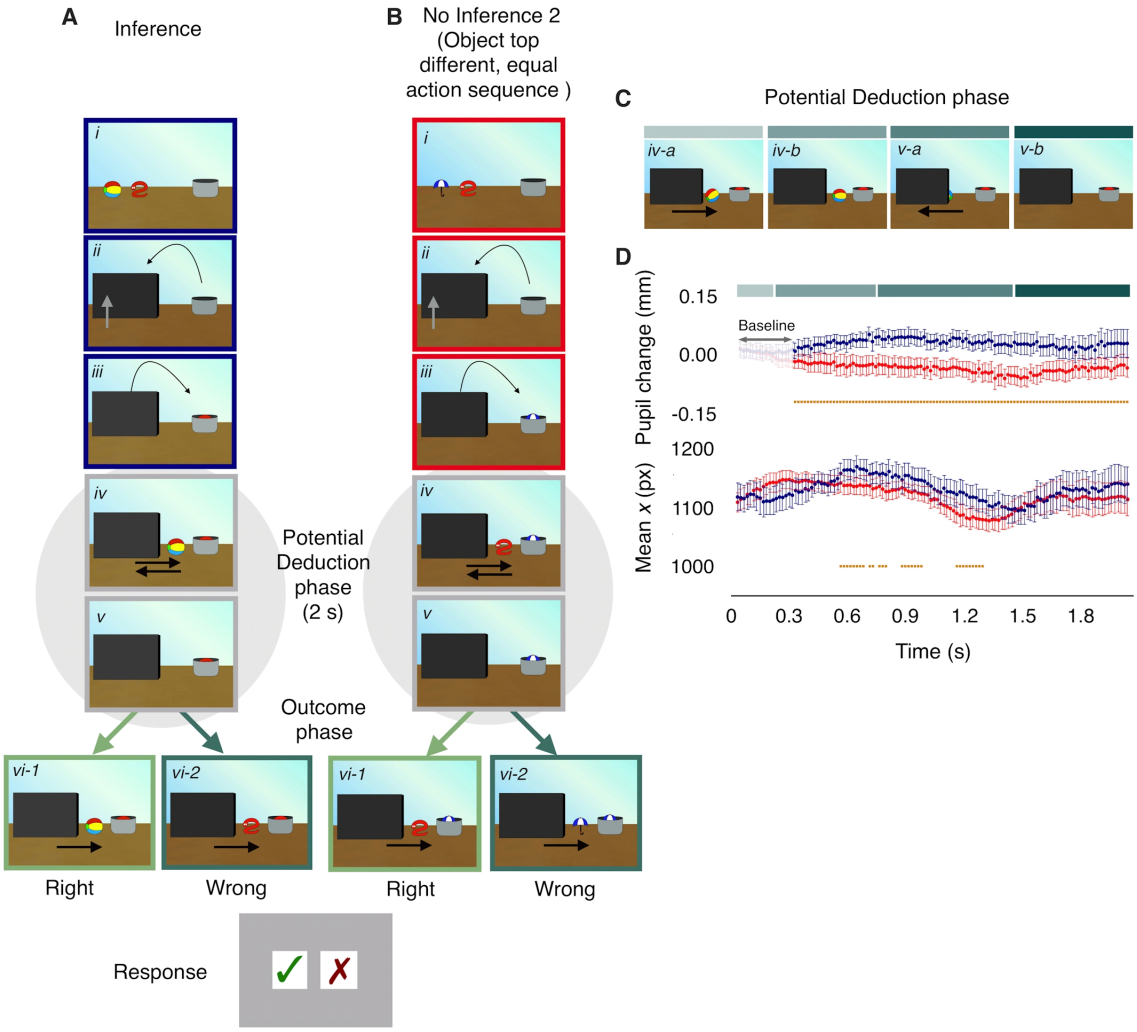

Fig. S6. Scene sequences for Inference and Type 2 No-Inference scenes of Experiment 7, and temporal analysis of the Potential Deduction phase (Inference triggered by the object exiting the occluder).
(**A**)In the Inference Scenes, the object in the cup was unknown, whereas in the (**B**) No-inference type 2 scenes it was always known. Scenes proceed identically in both conditions but in (B) objects never share the same upper part. Note that during the Potential Deduction phase the occluder remains and the object behind the occluder exits allowing to deduce the content of the cup. The Outcome phase (**vi**, identical for all scenes) are realized with the a second exiting of the object behind the occluder: either consistent with the logical deduction (**iv-1**) or inconsistent (**iv-2**). (**C**) Sub-phases of the Potential Deduction phase (**iv-v**), color-coded. (**D**) Temporal course of pupil dilations differences from baselines (mm) and Mean X gaze positions (px) (bottom). Statistical differences were used by computing a cluster mass test with 1000 sets of random

permutation generated at each sample of the 60 Hz data flow (one-tailed paired t-tests for pupil and two-tailed t-tests for mean-x position). Yellow dots indicate regions of differences in the two conditions. Adults had higher pupil dilation when looking at Inference than at No-Inference scenes. Error bars, 95% within-participants confidence intervals.

Table S1. Summary of the structure, functions and length of the familiarization movies in Experiments 1-6.

Experiment	Trial	Movie	Function	Length
1	1	movie 1	Familiarize with the occluder	1 x 12.31 s
1	2-3	movies 2-3 counterbalanced order	Familiarize with the cup and the occluder	2 x 15.81 s
1	4	movies 4	Familiarize with the test objects	1 x 13 s
2	1-2	movies 1-2 counterbalanced order	Familiarize with cup and occluder	2 x 13.48 s
2	3-4	movies 3-4 counterbalanced order	Familiarize with interaction of cup and occluder	2 x 11.73 s
2	5-6	movies 5-6 counterbalanced order	Familiarize with test trial structure	2 x 18.98 s
3 and 5	1-2	movies 1-2 counterbalanced order	Familiarize with cup and occluder	2 x 13.48 s
3 and 5	3-4	movies 3-4 counterbalanced order	Familiarize with interaction of cup and occluder	2 x 15.36 s
3 and 5	5-6	movies 5-6 counterbalanced order	Familiarize with the exit of objects from the occlude	2 x 14.79 s
3 and 5	7	movie 7	Familiarize with test objects	1 x 13 s
4 and 6	1-2	movies 1-2 counterbalanced order	Familiarize with cup and occluder	2 x 13.48 s
4 and 6	3-6	movies 3-6 random order	Familiarize with interaction of cup and occluder and with test objects.	4 x 13.86 s
4 and 6	7-8	7-8 counterbalanced order	Familiarize with test trial structure	2 x 17.98 s

Movie S1

Example of a movie from Experiment 1 (19-month-olds). Inference condition, cup content revealed, consistent outcome.

Movie S2

Example of a movie from Experiment 1 (19-month-olds). Inference condition, cup content revealed, inconsistent outcome.

Movie S3

Example of a movie from Experiment 4 (12-month-olds). Inference condition, cup content not revealed, consistent outcome.

Movie S4

Example of a movie from Experiment 4 (12-month-olds). Inference condition, cup content not revealed, inconsistent outcome.

Movie S5

Example of a movie from Experiment 6 (12-month-olds). No inference condition, cup content not revealed, consistent outcome.

Movie S6

Example of a movie from Experiment 6 (12-month-olds). No inference condition, cup content not revealed, inconsistent outcome.

Movie S7

Example of a movie from Experiment 7 (adults). Inference condition, cup content not revealed, wrong outcome.

Movie S8

Example of a movie from Experiment 7 (adults). No inference condition type 1, cup content not revealed, right outcome.

Movie S9

Example of a movie from Experiment 7 (adults). No inference condition type 2, cup content not revealed, right outcome.

References and Notes

1. M. Piattelli-Palmarini, *Language and Learning: The Debate Between Jean Piaget and Noam Chomsky* (Harvard Univ. Press, 1980).

2. E. S. Spelke, in *Language in Mind: Advances in the Study of Language and Thought*, D. Gentner, S. Goldin-Meadow, Eds. (MIT Press, 2003), pp. 277–311.

3. W. V. O. Quine, *Word and Object* (MIT Press, 1960).

4. D. C. Penn, K. J. Holyoak, D. J. Povinelli, Darwin's mistake: Explaining the discontinuity between human and nonhuman minds. *Behav. Brain Sci.* **31**, 109–130 (2008). Medline

5. S. Carey, *The Origin of Concepts* (Oxford Univ. Press, 2009).

6. J. R. Saffran, R. N. Aslin, E. L. Newport, Statistical learning by 8-month-old infants. *Science* **274**, 1926–1928 (1996). doi:10.1126/science.274.5294.1926 Medline

7. E. Téglás, V. Girotto, M. Gonzalez, L. L. Bonatti, Intuitions of probabilities shape expectations about the future at 12 months and beyond. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 19156–19159 (2007). doi:10.1073/pnas.0700271104 Medline

8. E. Téglás, E. Vul, V. Girotto, M. Gonzalez, J. B. Tenenbaum, L. L. Bonatti, Pure reasoning in 12-month-old infants as probabilistic inference. *Science* **332**, 1054–1059 (2011). doi:10.1126/science.1196404 Medline

9. E. Téglás, A. Ibanez-Lillo, A. Costa, L. L. Bonatti, Numerical representations and intuitions of probabilities at 12 months. *Dev. Sci.* **18**, 183–193 (2015). doi:10.1111/desc.12196 Medline

10. H. Gweon, J. B. Tenenbaum, L. E. Schulz, Infants consider both the sample and the sampling process in inductive generalization. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 9066–9071 (2010). doi:10.1073/pnas.1003095107 Medline

11. A. E. Stahl, L. Feigenson, Observing the unexpected enhances infants' learning and exploration. *Science* **348**, 91–94 (2015). doi:10.1126/science.aaa3799 Medline

12. H. Gweon, L. Schulz, 16-month-olds rationally infer causes of failed actions. *Science* **332**, 1524 (2011). doi:10.1126/science.1204493 Medline

13. A. Gopnik, The theory theory 2.0: Probabilistic models and cognitive development. *Child Dev. Perspect.* **5**, 161–163 (2011). doi:10.1111/j.1750-8606.2011.00179.x

14. S. T. Piantadosi, J. B. Tenenbaum, N. D. Goodman, The logical primitives of thought: Empirical foundations for compositional cognitive models. *Psychol. Rev.* **123**, 392–424 (2016). doi:10.1037/a0039980 Medline

15. S. J. Gershman, E. J. Horvitz, J. B. Tenenbaum, Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science* **349**, 273–278 (2015). doi:10.1126/science.aac6076 Medline

16. S. Carey, in *The Conceptual Mind: New Directions in the Study of Concepts*, E. Margolis, S. Laurence, Eds. (MIT Press, 2015), pp. 415–454.

17. S. Mody, S. Carey, The emergence of reasoning by the disjunctive syllogism in early childhood. *Cognition* **154**, 40–48 (2016). doi:10.1016/j.cognition.2016.05.012 Medline

18. J. Halberda, The development of a word-learning strategy. *Cognition* **87**, B23–B34 (2003). doi:10.1016/S0010-0277(02)00186-5 Medline

19. J. Halberda, Is this a dax which I see before me? Use of the logical argument disjunctive syllogism supports word-learning in children and adults. *Cognit. Psychol.* **53**, 310–344 (2006). doi:10.1016/j.cogpsych.2006.04.003 Medline

20. E. Téglás, L. L. Bonatti, Infants anticipate probabilistic but not deterministic outcomes. *Cognition* **157**, 227–236 (2016). doi:10.1016/j.cognition.2016.09.003 Medline

21. R. B. Lea, D. P. O'Brien, S. M. Fisch, I. A. Noveck, M. D. S. Braine, Predicting propositional logic inferences in text comprehension. *J. Mem. Lang.* **29**, 361–387 (1990). doi:10.1016/0749-596X(90)90005-K

22. C. Reverberi, D. Pischedda, M. Burigo, P. Cherubini, Deduction without awareness. *Acta Psychol.* **139**, 244–253 (2012). doi:10.1016/j.actpsy.2011.09.011 Medline

23. L. Bonatti, E. Frot, R. Zangl, J. Mehler, The human first hypothesis: Identification of conspecifics and individuation of objects in the young infant. *Cognit. Psychol.* **44**, 388–426 (2002). doi:10.1006/cogp.2002.0779 Medline

24. E. Maris, R. Oostenveld, Nonparametric statistical testing of EEG- and MEG-data. *J. Neurosci. Methods* **164**, 177–190 (2007). doi:10.1016/j.jneumeth.2007.03.024 Medline

25. J. R. Hochmann, L. Papeo, The invariance problem in infancy: A pupillometry study. *Psychol. Sci.* **25**, 2038–2046 (2014). doi:10.1177/0956797614547918 Medline

26. C. Reverberi, L. L. Bonatti, R. S. J. Frackowiak, E. Paulesu, P. Cherubini, E. Macaluso, Large scale brain activations predict reasoning profiles. *Neuroimage* **59**, 1752–1764 (2012). doi:10.1016/j.neuroimage.2011.08.027 Medline

27. G. Chierchia, *Logic in Grammar: Polarity, Free Choice, and Intervention* (MIT Press, 2013).

28. A. Papafragou, K. Cassidy, L. Gleitman, When we think about thinking: The acquisition of belief verbs. *Cognition* **105**, 125–165 (2007). doi:10.1016/j.cognition.2006.09.008 Medline

29. D. Kahneman, J. Beatty, Pupil diameter and load on memory. *Science* **154**, 1583–1585 (1966). doi:10.1126/science.154.3756.1583 Medline