nature human behaviour

Article

https://doi.org/10.1038/s41562-024-01965-x

Perceptual and conceptual novelty independently guide infant looking behaviour: a systematic review and meta-analysis

Received: 15 January 2024

Accepted: 23 July 2024

Published online: 14 October 2024

Check for updates

Linette Kunin[1], Sabrina H. Piccolo[2], Rebecca Saxe[1] & Shari Liu[3]

Human infants are born with their eyes open and an otherwise limited motor repertoire; thus, studies measuring infant looking are commonly used to investigate the developmental origins of perception and cognition. However, scholars have long expressed concerns about the reliability and interpretation of looking behaviours. We evaluated these concerns using a pre-registered (https://osf.io/jghc3), systematic meta-analysis of 76 published and unpublished studies of infants' early physical and psychological reasoning (total n = 1,899; 3- to 12-month-old infants; database search and call for unpublished studies conducted July to August 2022). We studied two effects in the same datasets: looking towards expected versus unexpected events (violation of expectation (VOE)) and looking towards visually familiar versus visually novel events (perceptual novelty (PN)). Most studies implemented methods to minimize the risk of bias (for example, ensuring that experimenters were naive to the conditions and reporting inter-rater reliability). There was mixed evidence about publication bias for the VOE effect. Most centrally to our research aims, we found that these two effects varied systematically—with roughly equal effect sizes (VOE, standardized mean difference 0.290 and 95% confidence interval (0.208, 0.372); PN, standardized mean difference 0.239 and 95% confidence interval (0.109, 0.369))—but independently, based on different predictors. Age predicted infants' looking responses to unexpected events, but not visually novel events. Habituation predicted infants' looking responses to visually novel events, but not unexpected events. From these findings, we suggest that conceptual and perceptual novelty independently influence infants' looking behaviour.

Studies of human infants offer a window to the developmental origins of the mind. For example, many experiments show that infants look longer at surprising physical outcomes (an object floats in midair) and surprising actions (an agent behaves inefficiently) relative to visually similar but expected events[1–3]. Longer looking at these surprising events, or the violation of expectation (VOE) effect, is taken as evidence for a hypothesized expectation held in infants' minds[4]: that unsupported objects fall and that agents tend to act efficiently[5–8]. However, looking measures have long been controversial. The robustness and nature of the VOE effect have been heavily debated, with some scholars claiming that behavioural effects in this literature are too noisy[9] or reducible to stimulus-driven confounds[10] and thus uninformative for studies about infant cognition[11]. Inspired by these concerns, in this Article we measure the size of the VOE effect and test whether the VOE effect is similar in size and source to looking behaviour driven by perceptual novelty (PN) across datasets from previous research.

A full list of affiliations appears at the end of the paper. e-mail: sliu199@jhu.edu

https://doi.org/10.1038/s41562-024-01965-x

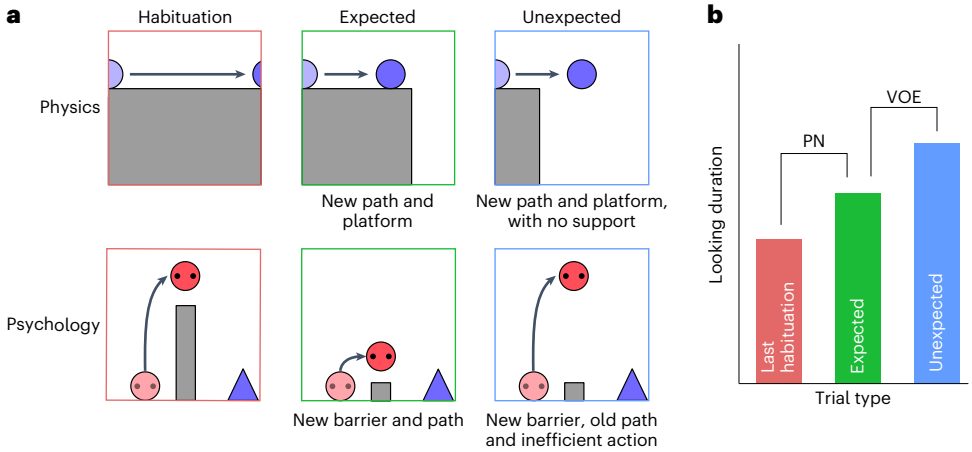


Fig. 1 | Sources of novelty in VOE studies. a, Schematics of stimuli from previous studies of early physical and psychological understanding. The physics task[7] (top row) tests whether infants expect objects to fall when unsupported. The psychology task[12] (bottom row) tests whether infants expect agents to act efficiently to reach their goals. In both tasks, after habituation or familiarization with one event, infants are shown two test events: an expected event and an unexpected event. Note that both event types introduce visual differences. In the current research, we ask whether changes that entail VOEs (lack of support or inefficient action) evoke similar or distinct looking behaviours when compared with changes in lower-level perceptual variables (a new path or new barrier). **b**, Barplot of idealized data depicting PN (difference between the expected test event and last habituation or familiarization event) and VOE (difference between unexpected and expected test events).

Logic and design of VOE studies

Consider the events in Fig. 1 from studies of infant cognition[7,12]. In a key surprising test event, infants see a ball roll off a platform and hover in midair or an agent jump inefficiently over an obstacle (Fig. 1a). These events are both visually novel and inconsistent with our expectations about objects and agents. When infants look longer at these events, which of these factors is driving that behaviour? To answer this question, developmental psychologists can first habituate infants to the agents and objects in the scene and then: (1) pit expectation- and visually driven novelty against each other; or (2) equate visual novelty across both expected and unexpected events. The expected outcomes in Fig. 1 are visually novel but do not violate the hypothesized expectation: the ball rolls across a new platform on a new trajectory and the agent jumps on a new but efficient path. The unexpected test events, in contrast, violate the hypothesized expectation but are visually familiar (the agent jumps in the same trajectory as during habituation, but now inefficiently) or are equally visually novel relative to the expected event (the ball rolls beyond the shorter platform on the same path as in the expected event). When infants look longer at the unexpected test events relative to the expected test events, this can serve as initial evidence that infants detect and prioritize the violation of the hypothesized expectation. Control conditions are then conducted to further test perceptual alternative explanations.

Open questions about the reliability and nature of VOE

Despite these methodological strategies, the VOE effect remains controversial. One critique is that infant research is conducted using small, underpowered samples[13] and thus the VOE effect could reflect an over-interpretation of noise. If true, with a high-powered test, researchers may not find a reliable effect at all, especially compared with more established effects, such as longer looking towards visually novel stimuli. A second critique is that longer looking to unexpected events is explained by lower-level features of the stimulus[9,11,14–16]. Infants attend to changes in colour[17], shape[18], brightness[19], spatial frequency[20] and motion[21]. Since the VOE method requires unexpected and expected events to be visually distinguishable, some researchers prefer the more parsimonious explanation that infant attention is driven primarily by a single mental process grounded in the visual features of the stimuli. Despite this debate, there has been no systematic analysis of the size or moderators of the VOE effect from previous studies of infant cognition (see refs. 22–24 for systematic analyses of other topics). Claims about the noisiness and source of VOE effects in this literature are often based on a small number of case studies and are rarely supported by large-scale quantitative evidence (but see refs. 25,26).

Overview of current research

In this Article, we contribute to this debate, concerning the reliability and nature of VOE, using the tools of meta-analysis and mega-analysis. Meta-analysis aggregates and quantifies condition-level effect sizes, whereas mega-analysis aggregates and quantifies effects from individual infants[27]. Using the framework and toolkit provided by MetaLab[28], we performed a systematic meta-analysis using studies on two classic infant cognition topics: inanimate objects and animate agents. Table 1 provides an overview of the papers that met our topic, methods and inclusion criteria. Figure 2 provides an overview of our data curation process and a summary of our final dataset (condition-level data from 76 studies and 1,899 infants aged 3–12 months (50.3% female) and infant-level data from 60 studies and 1,482 infants (51.2% female); many authors did not report demographic information, but these studies probably follow the past trends of the field[29], focusing on predominantly White populations from North America and Western Europe). We chose to focus on studies from the first year of life because controversy regarding infant looking is often centred around young infants[9]. To focus our investigation on datasets that measured both looking towards unexpected stimuli and looking towards visually novel stimuli, our primary analyses included only studies labelled by the authors as experimental conditions (76 studies and 1,899 infants; see Supplementary Information for additional analyses on control conditions). We acquired original datasets, including data about the age of individual infants and the order in which they saw the test events, from 60 studies and 1,482 infants; these were the focus of our mega-analyses. All studies included in the analyses used an infant-controlled design: on each trial, infants' looking was monitored and the trial was terminated when infants looked away for a set duration (typically 2 s). In most studies, the trial could also be terminated after infants looked for some maximum duration or after a fixed number of seconds had passed (typically 30–120 s). In all studies, looking duration in a trial was defined as the number of seconds for which infants looked at the stimuli before the trial

Table 1 | Information on papers included in our analyses

Study ID	Short citation	Infant-level data available	Number of studies	Number of infants
Biro_2007	Biro et al.[80]	No	3	126
Brandone_2009	Brandone and Wellman[81]	Yes	8	182
Choi_2018	Choi et al.[82]	Yes	3	48
Chuey_2021	Chuey et al.[83]	Yes	1	30
Gerson_2014a	Gerson and Woodward[84]	Yes	3	72
Gerson_2014b	Gerson and Woodward[85]	Yes	3	90
Hernik_2012	Hernik and Southgate[86]	No	3	48
Hespos_2009	Hespos et al.[87]	Yes	2	31
Jackson_2022	Jackson and Sirois[16]	No	1	24
Lakusta_2015	Lakusta and Carey[88]	No	3	60
Liu_2017a	Liu and Spelke[12]	Yes	3	60
Liu_2017b	Liu et al.[48]	Yes	3	80
Liu_2019	Liu et al.[89]	Yes	7	152
Liu_2022	Liu et al.[90]	Yes	3	102
Liu_unpublisheda	S. Liu et al. (unpublished-a), Liu et al.[91]	Yes	2	80
Liu_unpublishedb	S. Liu et al. (unpublished-b), Liu et al.[91]	Yes	1	26
Liu_unpublishedc	S. Liu and E. S. Spelke (unpublished-a), Liu and Spelke[92]	Yes	1	37
Liu_unpublishedd	S. Liu and E. S. Spelke (unpublished-b), Liu and Spelke[93]	Yes	1	31
Luo_2005a	Luo and Baillargeon[94]	Yes	4	48
Luo_2005b	Luo and Baillargeon[95]	Yes	2	16
Luo_2009a	Luo and Johnson[96]	Yes	5	60
Luo_2009b	Luo et al.[97]	Yes	4	64
Luo_2010	Luo[98]	Yes	4	40
Luo_2011	Luo[99]	Yes	3	36
Martin_2017	Martin et al.[100]	Yes	8	160
Olofson_2011	Olofson and Baldwin[101]	No	2	32
Powell_unpublished	L. Powell, A. Schachner and E. Spelke (unpublished)	Yes	1	60
Schlottmann_2012	Schlottmann et al.[102]	No	1	56
Skerry_2013	Skerry et al.[103]	Yes	5	112
Spaepen_2007	Spaepen and Spelke[104]	No	5	84
Stojnic_2023	Stojnić et al.[46]	Yes	8	336
Thoermer_2013	Thoermer et al.[105]	Yes (no data on age provided)	1	43
Woo_2021	Woo et al.[106]	Yes	3	68

Each study contributed average looking durations (as well as individual infants' looking durations in 89 studies) for the last trial before the test, the first expected test trial and the first unexpected test trial; a subset of these studies also contributed data from individual infants. Note that this table includes studies that were experimental conditions (in which VOE and PN could both be measured) and studies that were negative control conditions (which by hypothesis should not evoke a VOE effect). The analyses in the main text focus on data from experimental conditions (see Supplementary Information for further analyses of data from studies under control conditions).

ended. Thirty-eight studies employed a habituation procedure (wherein the number of trials before the test events varied as a function of how quickly infants decreased their looking) and 38 studies employed a familiarization procedure (with a fixed number of trials before the test events). The final meta-analytic and mega-analytic datasets are openly available at https://osf.io/b59km/. See Methods for additional details.

$$\mathrm{VOE} = \mathrm{looking}_{\mathrm{unexpected}} - \mathrm{looking}_{\mathrm{expected}} \quad (1)$$

$$\mathrm{PN} = \mathrm{looking}_{\mathrm{expected}} - \mathrm{looking}_{\mathrm{last\ habituation}} \quad (2)$$

For each study, we defined two looking time effects. First, we defined the violation-of-expectation (VOE) effect as the difference between infants' looking on the first expected and first unexpected test trials (equation (1)). This standard definition in the literature is probably a conservative estimate since expected events are also visually novel—and sometimes more visually novel than the unexpected event—which could lead to a smaller looking preference for the unexpected event. Second, we defined a measure of responses to perceptual novelty (PN) in the same datasets: the difference between the last habituation or familiarization trial and the first expected test trial, which contained a visually novel but conceptually expected outcome (equation (2); see Extended Data Fig. 1).

https://doi.org/10.1038/s41562-024-01965-x

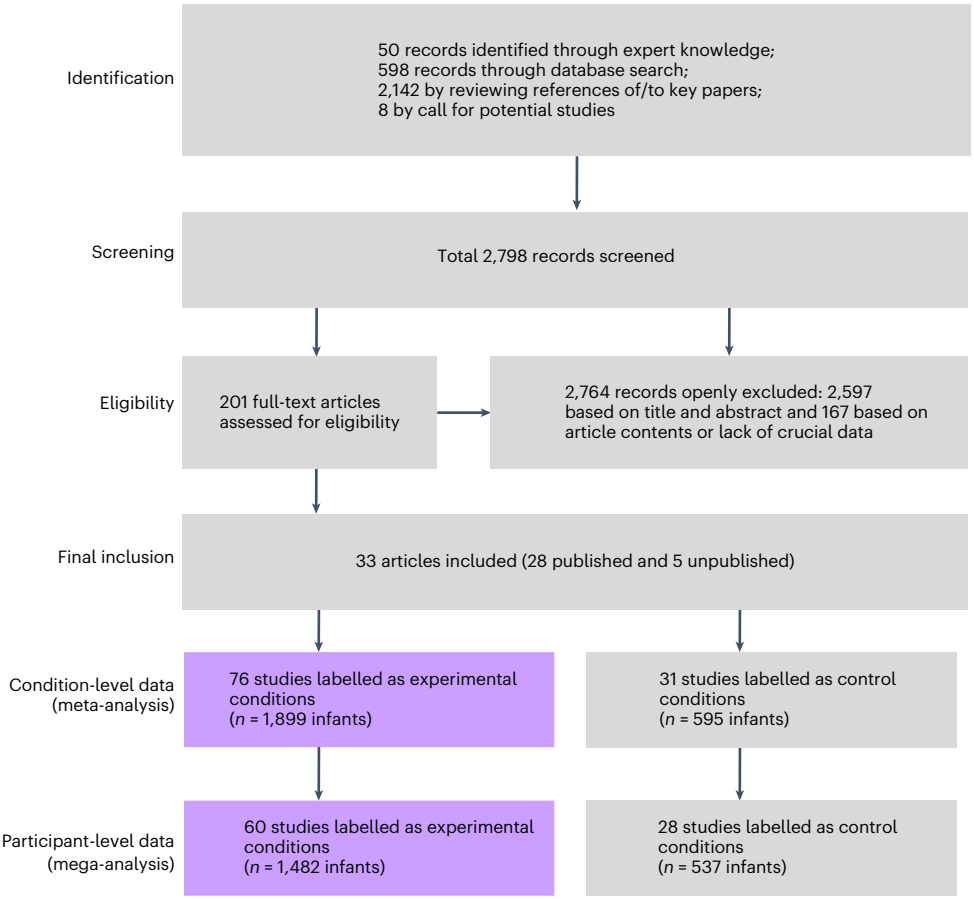


Fig. 2 | Condensed PRISMA diagram for the current research. This figure summarizes the methods for identifying potential studies and screening them against our eligibility criteria. The final datasets included in our primary analyses are shown in purple (see Extended Data Fig. 7 for a full PRISMA diagram and Supplementary Information for the results from studies serving as negative controls). We excluded one outlier paper (two studies) that passed our screening process, but skewed some of our supplemental meta-analytic results due to its extremely low variance relative to the other studies. Our primary conclusions hold regardless of whether this paper is included (see Supplementary Information for details).

For all studies, these three trials appeared consecutively (for example, last habituation, first expected, first unexpected) and the order of the test events (expected first or unexpected first) was counterbalanced across infants. By measuring the VOE and PN effects in the same experiments, we ensured that all methodological variables were identical across effects: infants were habituated or familiarized with the same stimuli before encountering test trials involving conceptually novel (unexpected) and perceptually novel (visually new) events. Under this design, infants in the first phase of the experiment were shown both a set of visual features (for example, the path of motion from Fig. 1a) and a set of conceptually relevant features (for example, there was an agent, navigating around a barrier towards a goal object). Then, at test, each of these features was changed in turn, allowing us to measure infants' looking behaviour based on each change. This approach allowed us to estimate the size of the VOE and PN effects using the same statistical models, controlling for the size of the other effect, within studies and infants. Lastly, analysing data aggregated from many studies, articles and laboratories allowed us to draw general conclusions about this literature.

First, to address the concern that VOE results from the overinterpretation of small noisy samples, we asked whether this effect is reliably larger than zero in a well-powered test, in comparison to the less controversial PN effect. We also evaluated evidence for publication bias in this literature, which could lead to an overestimate of the VOE effect, and conducted a power analysis to estimate the number of infant participants needed to detect both effects in a new sample, which we report in the Supplementary Information.

Second, we tested the claim that longer looking during VOE studies is primarily driven by low-level visual features in the stimuli. To do so, we compared the moderators of the VOE and PN effects. Beyond trial type, which was a within-participants predictor required to calculate the PN and VOE effects, we also considered a pre-registered set of between-participants predictors (Table 2). These included features of the infants that were tested, such as infant age, and features of the experimental design, such as whether infants were habituated or familiarized before the test events and the domain of knowledge the experiment was testing. Some predictors were chosen because they are theoretically interesting (for example, age and domain). Others were chosen because they are likely to explain variance in infant looking (for example, trial order). We pre-registered two alternative hypotheses. If the VOE effect is reducible to—and indistinguishable from—a response to PN, we should find that the same factors moderate the size of both effects. If however these effects reflect distinct motives of infant looking, we could find that different moderating factors predict the size of each effect. We also made several predictions about the size of the PN effect: that younger infants would show a bigger PN effect than older infants due to differences in endogenous control of attention; that negative control studies would report bigger PN effects than studies hypothesized to evoke the VOE effect, because they only included visual changes and not unexpected events (see Supplementary Information for the results from negative control studies); and that habituation studies would report bigger PN effects than familiarization studies because they provide infants with more time to encode the stimuli before the test events.

Table 2 | Overview of fixed effects

Name	Description	Type	Values
trial_type	The trial from which looking times were collected	Factor	'Last_train' (last familiarization or habituation trial), 'expected' (first expected test trial) or 'unexpected' (first unexpected test trial)
mean_age	Average age of infants per study (d)	Numeric	See infant_age, which is the same moderator, but includes data from individual infants
equal_per_nov	Relative to familiarization or habituation, whether the expected and unexpected test events were equally perceptually novel (as in the top row of Fig. 1a), or the expected test event was more perceptually novel than the unexpected test event (as in the bottom row of Fig. 1a)	Factor	'Yes' or 'no'
exposure_phase	Whether infants were habituated or familiarized before the test events	Factor	'Habituation' (variable number of trials across infants) or 'familiarization' (fixed number of trials)
domain	The type of knowledge the experiment tested	Factor	'Physics' or 'psychology' (if potentially both, we chose the domain most emphasized by the authors of the paper)
stim_loop	Whether the stimulus repeated on each trial until infants looked away (or the stimulus was shown only once)	Factor	'Yes' (repeated) or 'no' (shown once)
infant_age	Age of individual infants (d)	Numeric	See mean_age, which is the same moderator, but averaged across infants per study
order	Which test event an infant was assigned to see first	Factor	'Expected' or 'unexpected'
exp_or_control	Whether the study was an experimental study (testing the hypothesized expectation) or a negative control (testing an alternative explanation for the results of the experimental study and predicting a null effect)	Factor	'Experimental' or 'control'

Overall, we found that infant looking is driven by perceptual novelty (PN), or visual changes to the stimulus and conceptual novelty (VOE), or the unexpectedness of that stimulus to a similar degree. Additionally, these effects were moderated by different predictors: the PN effect was bigger for studies that used habituation and for individual infants who showed stronger habituation effects (with no such effect for VOE); and the VOE effect was smaller in studies testing older infants (with no such effect for PN). From these findings, we suggest that stimulus- and expectation-driven novelty independently guide infant looking behaviour.

Results

All of the results and figures presented in the main text exclude one outlier paper[30], which heavily skewed some of the supplementary meta-analytic results due to its extremely low variance relative to the other studies. The results including this study are presented in the Supplementary Information for full transparency. All of our primary results hold, regardless of whether this paper is excluded or included.

Comparing the magnitudes of VOE and PN effects

How much do unexpected stimuli drive infants' looking behaviour (VOE, the primary effect under discussion) relative to visually novel stimuli (PN, a less controversial effect that looking time studies were invented to measure)? In a confirmatory analysis across 76 studies and 1,899 infants, we found that both the PN and VOE effects were significantly greater than 0 (PN mean difference = 1.700 s ($z = 7.712$; $P < 0.001$; two-tailed test; effect size (standardized mean difference (SMD)) = 0.239; 95% confidence interval (CI) = (1.268, 2.132)); VOE mean difference = 2.128 s ($z = 8.575$; $P < 0.001$; two-tailed test; effect size (SMD) = 0.290; 95% CI = (1.641, 2.614)); Fig. 3). We then directly compared the size of these two effects in standardized mean units; these two effect sizes were not significantly different from each other ($z = 1.305$; $P = 0.192$; two-tailed test; effect size = 0.061 (SMD); 95% CI = (−0.031, 0.152)) and the Bayes factor (0.190) strongly favoured the hypothesis that there was no difference between the two effects. See Fig. 3 for an aggregated summary of both effects and Fig. 4 and Extended Data Fig. 1 for information on the distribution of effect sizes across studies.

Publication bias. One challenge in using meta-analysis to estimate effect sizes is that the publication process may be biased towards publishing significant or inflated effects. In this literature, is there evidence for the preferential publication of studies with significant effects or imprecise studies with strongly positive effects? In exploratory analyses, we found mixed evidence for publication bias for the VOE effect. Using Egger's test for funnel plot asymmetry (see Extended Data Fig. 2), we found that the distribution of effect sizes relative to study precision was not asymmetrical for the PN effect ($b = 0.049$; 95% CI = (−0.512, 0.610); $z = 0.682$; $P = 0.495$), but was asymmetrical for the VOE effect ($b = -0.488$; 95% CI = (−0.804, −0.171); $z = 4.912$; $P < 0.001$). This suggests some degree of publication bias for the VOE effect and no evidence for publication bias for the PN effect (which was not directly studied but was incidentally measured in this literature). In contrast, using selection models, we found no evidence for publication bias for either effect (VOE: X^2(d.f.) = 1.002(1), $P = 0.317$; PN: X^2(d.f.) = 0.934(1), $P = 0.334$; likelihood ratio test).

What are the implications of these publication bias results for the current and subsequent analyses? First, it is plausible that we overestimated the size of the VOE effect because imprecise studies that show small or negative VOE effects (by Egger's test) or non-significant results (by selection models) could be missing from our dataset. Although we found mixed evidence for publication bias, we computed an adjusted estimate for the VOE effect using trim-and-fill[31] and selection models[32]. Using the trim-and-fill method, the adjusted effect size for VOE ($z = 3.582$; $P < 0.001$; two-tailed test; effect size (SMD) = 0.166; 95% CI = (0.075, 0.257); $I^2 = 34.96\%$, or moderate heterogeneity, before trim and fill) was still similar in size (with overlapping CIs) to the PN effect for which we found no evidence for publication bias ($z = 3.606$; $P < 0.001$; two-tailed test; effect size (SMD) = 0.239; 95% CI = (0.109, 0.369)). Because we found evidence for publication bias using the trim-and-fill method, we also examined the adjusted estimate for the VOE effect from the selection model. The adjusted VOE effect was still above zero, with CIs that overlapped with the PN effect ($z = 3.478$; $P < 0.001$; two-tailed test; effect size (SMD) = 0.228; 95% CI = (0.100, 0.357)). Overall, our results show that the VOE effect is robustly above 0 and as large as the PN effect after accounting for the possibility of publication

https://doi.org/10.1038/s41562-024-01965-x

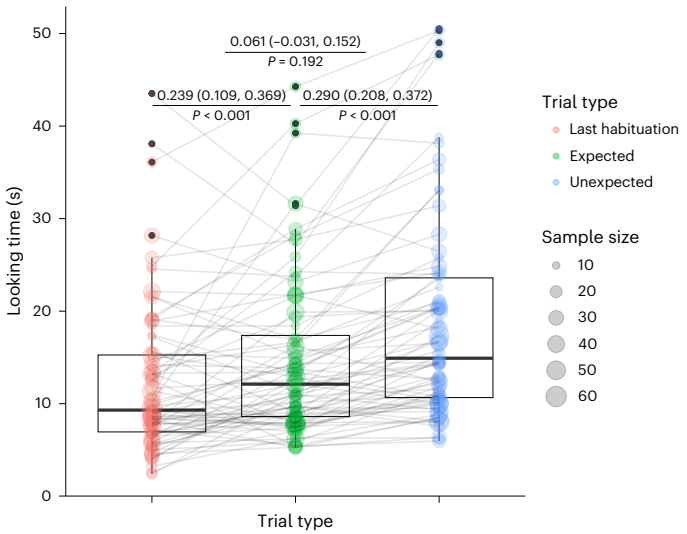


Fig. 3 | VOE and PN effects across previous literature. Each point represents the average looking time towards each of the three trial types (last habituation or familiarization, first expected, and first unexpected) for the 76 studies included in our confirmatory meta-analysis (n = 76 studies with data on 1,899 3- to 12-month-old infants). Point sizes indicate sample size (number of infants) per study. Lines connect data from the same studies. In each boxplot, the centre of the box indicates the median, the bounds of the box correspond to the 25th and 75th percentiles (the interquartile range (IQR)) and the whiskers extend to the minima and maxima (up to 1.5× the IQR from the 25th and 75th percentiles). Data beyond the ends of the whiskers are plotted in dark grey. The quartiles are unweighted (do not take into account differences in the sample sizes or variances across studies). The sizes of the PN and VOE effects—estimated using an intercept-only random effects meta-analysis with no fixed effects and study ID as a random intercept—are indicated in SMDs, along with 95% CIs (in parentheses) and two-tailed P values. Effect sizes of the difference between effect sizes are listed in the same units and came from a mixed effects model with effect as an outcome measure, effect type (PN versus VOE) as a fixed effect and study ID as a random intercept. Both PN and VOE effects were significantly greater than 0 (PN mean difference = 1.700 s (z = 7.712; P < 0.001; two-tailed test; effect size in SMD = 0.239; 95% CI = (1.268, 2.132)); VOE mean difference = 2.128 s (z = 8.575; P < 0.001; two-tailed test; effect size in SMD = 0.290; 95% CI = (1.641, 2.614))). These two effect sizes were not significantly different from each other (z = 1.305; P = 0.192; two-tailed test; effect size = 0.061 SMDs; 95% CI = (−0.031, 0.152)) and the Bayes factor (0.190) strongly favoured the hypothesis that there is no difference between the two effects.

bias. Since publication bias from these studies should primarily hinge on the size and significance of the VOE effect, this is less likely to affect moderators for the VOE and PN effects; thus, in subsequent moderator analyses, we did not test or correct for publication bias.

Additive versus multiplicative form. Are VOE and PN effects additive (for example, looking at unexpected events is ~2 s longer than looking at expected events) or multiplicative (~1.25× longer)? In exploratory analyses, we found that a model that expressed these study-level effects as multiplicative (log ratio of means) fit the data better than a model that expressed these effects as additive (SMD). This was true for both the VOE effect (log ratio of means: Akaike information criterion (AIC) = 19.872; Bayesian information criterion (BIC) = 24.533; versus SMD: AIC = 75.039; BIC = 79.700) and the PN effect (log ratio of means: AIC = 104.456; BIC = 109.117; versus SMD: AIC = 138.095; BIC = 142.756). In line with previous work[33–35], both effects are best conceived as ratios (PN ratio = 1.228 (z = 3.805; P < 0.001; two-tailed test; 95% CI = (1.105, 1.365)); VOE ratio = 1.253 (z = 7.260; P < 0.001; two-tailed test; 95% CI = (1.179, 1.331))) rather than differences. In subsequent moderator analyses over study-level data, we chose to continue modelling the means and sampling variances of looking time to each trial type, as originally pre-registered, which enabled us to straightforwardly model both effects simultaneously.

Comparing the moderators of VOE and PN effects

Next, in a series of pre-registered exploratory analyses, we asked whether similar or different study- and participant-level moderators predict looking towards unexpected and visually novel stimuli.

Modelling each effect separately, we found that these two looking behaviours were moderated by distinct predictors (Fig. 5). For the PN effect, the only significant moderator was whether studies used habituation or familiarization. Studies that habituated infants (38 studies) rather than familiarizing them for a fixed number of trials (38 studies) evoked a greater PN effect (z = 5.793; P = <0.001; two-tailed test; SMD = 0.338; 95% CI = (0.223, 0.452)). For the VOE effect, the only significant moderator was infant age. Studies on older infants reported a smaller VOE effect than studies on younger infants (z = −2.185; P = 0.029; two-tailed test; SMD = −0.100; 95% CI = (−0.190, −0.010)). In contrast, exposure phase did not significantly moderate the VOE effect (z = −1.404; P = 0.160; two-tailed test; SMD = −0.062; 95% CI = (−0.148, 0.024)) and age did not significantly moderate the PN effect (z = 1.462; P = 0.144; two-tailed test; SMD = 0.088; 95% CI = (−0.030, 0.205)). See Extended Data Figs. 3 and 4 for the effects of exposure phase and infant age on looking behaviour for each trial type. Descriptively, habituation studies reported shorter looking times than familiarization studies for all three trial types, but this effect was particularly strong for the last habituation or familiarization trial (thus selectively impacting the PN effect). Descriptively, younger infants looked longer than older infants on all three trial types, but this effect was particularly strong for unexpected events (thus selectively impacting the VOE effect). Our moderator analysis also revealed some potentially interesting negative results. We found no evidence that VOE effects differed across the domains of physical and psychological reasoning (z = 0.914; P = 0.361; two-tailed test; SMD = 0.061; 95% CI = (−0.070, 0.192)) or depended on whether the unexpected event was made to compete against an expected test event that was more visually novel (z = −0.655; P = 0.513; two-tailed test; SMD = −0.035; 95% CI = (−0.139, 0.070)).

Because the PN and VOE effects share a common input value (looking towards the expected test event), we next tested whether exposure phase and infant age differentially explained the VOE and PN effects when both effects were estimated in the same model[36] (Fig. 6). For both moderators, we found a significant interaction between that moderator and trial type (exposure phase × trial type: X^2(d.f.) = 134.057(2), P < 0.001); infant age × trial type: X^2(d.f.) = 18.039(2), P < 0.001); likelihood ratio test). Contrasts extracted from these models confirmed the findings that familiarization studies reported smaller PN effects than habituation studies (estimate = −4.568 s; z = −10.136; P < 0.001; two-tailed test; 95% CI = (−5.452, −3.685)), with no differences for VOE (estimate = −0.196 s; z = −0.394; P = 0.694; two-tailed test; 95% CI = (−1.169, 0.778)), and that infant age was associated with the size of the VOE effect (estimate = −1.323 s; z = −4.223; P < 0.001; two-tailed test; 95% CI = (−1.937, −0.709)) but not the PN effect (estimate = 0.449 s; z = 1.755; P = 0.079; two-tailed test; 95% CI = (−0.053, 0.951)). In summary, the VOE effect and PN effect were moderated by distinct predictors.

Results from individual infants. Do these study-level results hold up in analyses of individual infants? In pre-registered exploratory analyses, we repeated the meta-analyses for infant-level data (n = 1,482 from 60 studies) with a dependent measure (log-transformed looking time) that captures the multiplicative nature of the VOE and PN effects. For instance, two infants who show a looking preference ratio of 2 (2 and 4 s for infant one and 4 and 8 s for infant two) would show equivalent looking preferences in log seconds (log[4] − log[2] = log[8] − log[4] = 0.693). The mega-analytic results confirmed the meta-analytic results: the VOE and PN effects were not significantly different in size, but were

https://doi.org/10.1038/s41562-024-01965-x

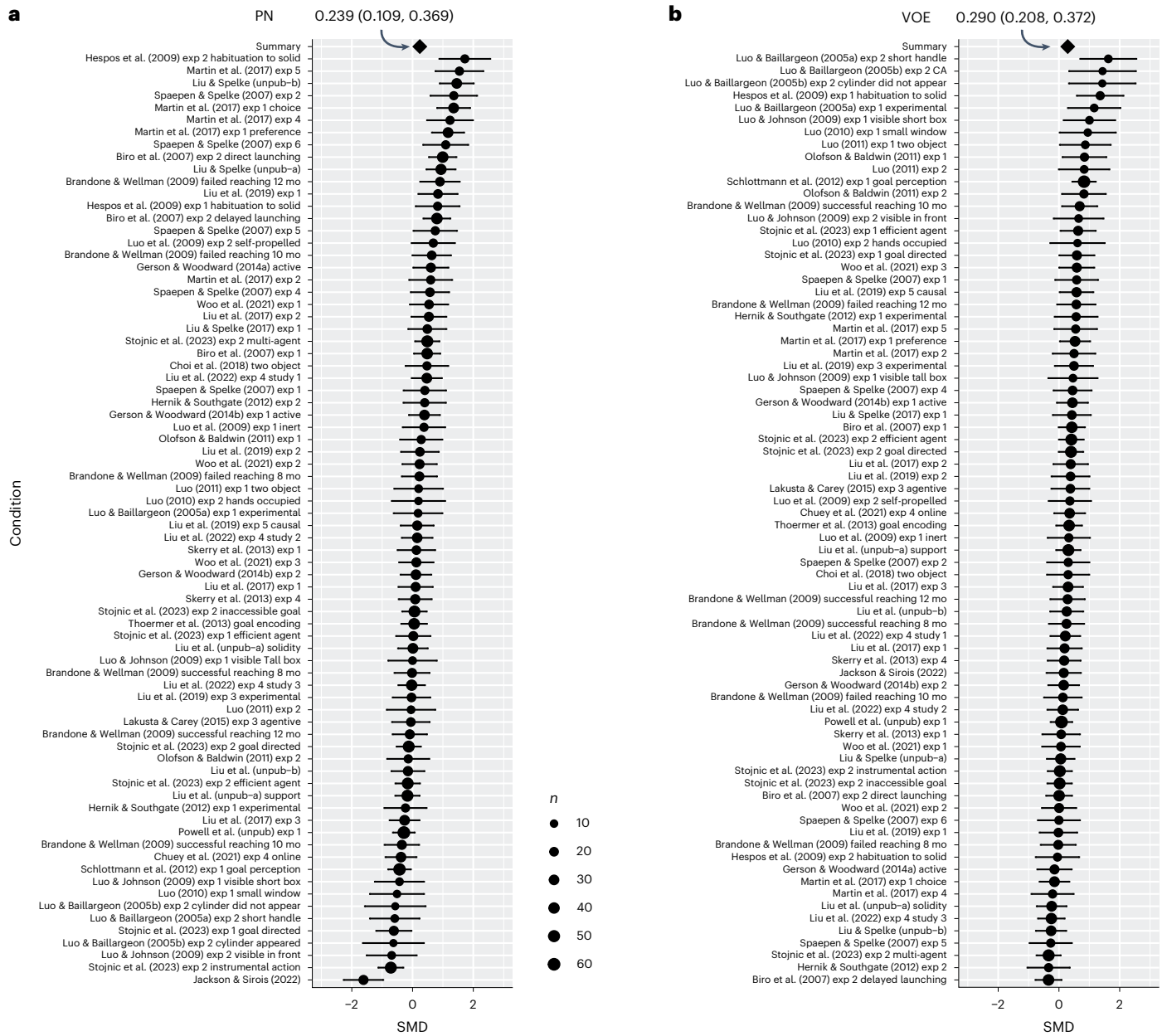


Fig. 4 | Forest plots of PN and VOE effects. a,**b**, Within each plot, estimates of the (**a**) PN and (**b**) VOE effects from 76 studies (n = 1,899; 3- to 12-month-old infants) are sorted by effect size (SMDs). Point sizes indicate sample sizes. The error bars represent 95% CIs around the effect size per study. The meta-analytic estimate of and 95% CI around each effect, indicated at the top of each panel, is the intercept of two separate mixed effects meta-analyses, with no fixed effects and a random intercept per study. exp, experiment; mo, months.

moderated by infant age and habituation, respectively (see Supplementary Information for details).

Data from individual infants allowed us to explore the differential effects of habituation on PN and VOE. Do habituation studies find bigger PN effects than familiarization studies because habituation selectively affects looking towards visually novel versus visually familiar stimuli, or could this finding be explained by other systematic differences between habituation and familiarization studies? We reasoned that if habituation genuinely affects PN and not VOE, individual differences in habituation rate should predict the size of the PN but not the VOE effect across both habituation and familiarization studies. We found evidence supporting this prediction. In habituation studies (499 infants; 22 studies), we found an interaction between trial type and the number of habituation trials infants underwent (likelihood ratio test: X^2(d.f.) = 19.6(2), P < 0.001): infants who habituated more steeply (underwent fewer habituation trials) showed a bigger PN effect (estimate = −0.06 log seconds; t = −3.662; d.f. = 993; P < 0.001; two-tailed test; standardized beta (β) = −0.175; 95% CI = (−0.268, −0.081)), but this did not predict the size of the VOE effect (estimate = −0.005 log seconds; t = −0.325; d.f. = 994; P = 0.745; two-tailed test; β = −0.016; 95% CI = (−0.109, 0.078)). In familiarization studies (603 infants; 21 studies), infants who would have met a standard habituation criterion (looking for a summed duration on the last three trials that was 50% or less than the summed duration on the first three trials[37]) showed a bigger PN effect than infants who did not (estimate = 0.531 log seconds; t = 7.25; d.f. = 1,187.363; P < 0.001; two-tailed test; β = 0.583; 95% CI = (0.425, 0.741)); this factor did not predict the size of the VOE effect (estimate = −0.008 log seconds; t = −0.114; d.f. = 1,184.622; P = 0.909; two-tailed test; β = −0.009; 95% CI = (−0.168, 0.150); interaction between trial type and habituation status: X^2(d.f.) = 67.3(2),

https://doi.org/10.1038/s41562-024-01965-x

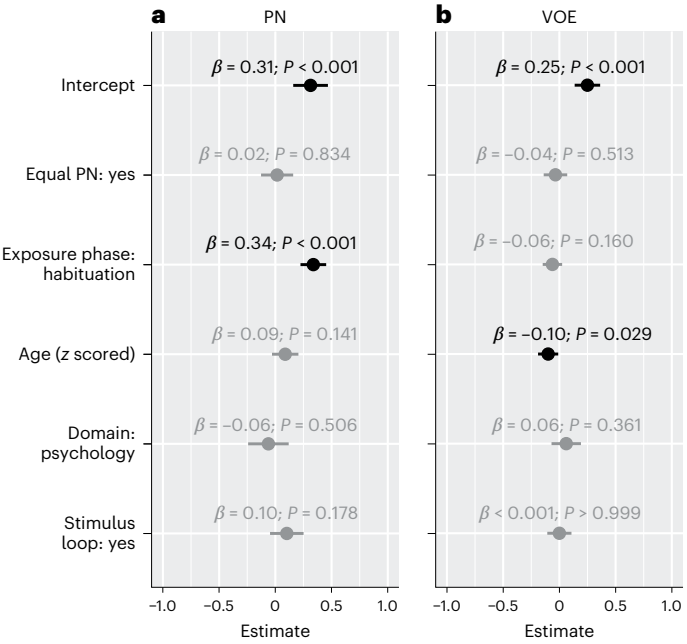


Fig. 5 | Moderators of the PN and VOE effects. a,**b**, Estimates of each moderator's effect (in SMDs, with 95% CIs) on the PN (**a**) and VOE effects (**b**) (n = 76 studies; 1,899 3- to 12-month-old infants) relative to the intercept (grand mean). The estimates came from two separate random effects meta-analyses (one for PN and one for VOE) with these moderators added as fixed effects and a random intercept for each study. Point estimates for each effect and two-tailed P values are listed with each moderator. From top to bottom, after the intercept, the fixed effects are: whether the expected and unexpected events were equally perceptually novel relative to familiarization (yes) or the expected event was more perceptually novel than the unexpected event (no); the exposure phase of the experiment (habituation versus familiarization); the average age of the infants (z scored across studies); the task domain (psychology versus physics); and whether the stimuli in the experiment were played on a loop (yes versus no). See Table 2 for details. Effects that passed our significance threshold (P < 0.05; two tailed) are indicated in black (otherwise, in grey). Studies that habituated infants (38 studies) rather than familiarizing them for a fixed number of trials (38 studies) evoked a greater PN effect (z = 5.793; P = <0.001; two-tailed test; SMD = 0.338; 95% CI = (0.223, 0.452)). Studies on older infants relative to younger infants reported a smaller VOE effect (z = −2.185; P = 0.029; two-tailed test; SMD = −0.100; 95% CI = (−0.190, −0.010)). See Supplementary Tables 2 and 3 for full results.

P < 0.001); likelihood ratio test). See Extended Data Figs. 5 and 6 and Methods for further details about these analyses. In summary, individual differences in habituation rate selectively moderated the PN effect across all experiments: infants who habituated faster (during habituation studies) or at all (during familiarization studies) recovered more attention to visually novel events but not unexpected events. We take this as our strongest piece of evidence that stimulus- and expectation-driven novelty independently contribute to infant looking behaviour.

Discussion

Infant looking is the most common behavioural measure used to study the developmental origins of perception and cognition in the first year of life[38,39]. In VOE experiments, longer looking towards unexpected versus expected events is often interpreted as evidence for infants' expectations about the social and physical world. Yet, many researchers have voiced concerns that the VOE effect is too noisy to interpret and boils down to a more basic process of stimulus-driven attention.

Here, we evaluated these critiques by curating and analysing a large dataset of experiments from previous literature. First, contrary to suggestions that VOE effects are too noisy and unreliable to trust[9], we found that showing infants something unexpected increased their looking as much as showing infants something visually novel, even after accounting for potential publication bias. Relatedly, contrary to claims that VOE effects appear in either direction (longer looking at unexpected or expected stimuli), thus rendering them difficult to interpret[11,40], we found no evidence for a bimodal distribution of either the PN or VOE effect (see Extended Data Fig. 1). Instead, most experiments observed effects that were numerically larger than 0 (62/76 for VOE and 49/76 for PN). Infants look longer at visual stimuli that are conceptually unexpected, not just those that are perceptually novel.

We then addressed a second critique that infant looking behaviour is primarily driven by low-level features in the stimuli. Under this hypothesis, VOE experiments measure the same PN effect twice: once based on low-level visual differences between the expected event and the familiarization events; and a second time based on low-level visual differences between the unexpected event and the expected event. Yet, when we studied the infant- and study-level moderators of both effects, we found that the VOE and PN effects were moderated by distinct predictors. The PN effect was bigger when infants were habituated instead of familiarized and the VOE effect was bigger in younger infants. In summary, these two looking behaviours—longer looking towards unexpected events and longer looking towards visually novel events—vary independently, but systematically, within the same experiments. Therefore, processing of conceptually unexpected information and processing of visually novel information are distinct—or at least not identical—processes in infant minds.

Habituation to familiar stimuli and orientation towards novel stimuli has long been identified as a basic behaviour of dynamic living systems, from slime moulds to people[40–44]. When infants decrease their looking in response to repeated stimuli, what are they habituating to? Here, we found that infants who habituated at all (in familiarization studies) or faster (in habituation studies) showed greater dishabituation to visually novel stimuli. In contrast, infants' habituation rate did not predict looking responses to unexpected stimuli. Therefore, one implication of our findings is that behavioural habituation in infant VOE experiments reflects low-level visual encoding, rather than a higher-level process of building expectations about the agents and objects in the stimuli. However, infants can clearly learn to extract higher-level statistical regularities over the time course of an experiment[45]; it is an open question whether the habituation rate in those experiments is predictive of this learning.

Computational models of cognition and development aspire to provide explicit and formal accounts of the origins of the mind. These models tend to prioritize low-level prediction (for example, the next frame in a video stimulus[46,47]) or higher-level prediction (for example, the probability of encountering the current stimulus, given a mental model of the situation[48,49]) and sometimes both (for example, next frame prediction over object representations[47,50,51]). Our results suggest that a full formal account of the infant mind should respond to both perceptually novel and conceptually unexpected events, but that these two signals of novelty should be modelled separately.

The current work has limitations. First, we studied infants' basic expectations about objects (for example, solidity and permanence) and agents (for example, goal-directed, efficient action) because these are two of the oldest topics in infant cognition, with a relatively large number of studies devoted to them. As a result, our claims only apply to these studies. However, our methods could be adapted to study any topic across the developmental and behavioural sciences.

Second, due to the retrospective nature of meta-analysis, the studies included in our analyses probably differed along other dimensions beyond the ones we chose to study. As a result, it is difficult to interpret most of the between-study moderator effects. For example, we predicted that the PN effect would be larger in younger infants because they have poorer endogenous control over their attention, but the null result we obtained is difficult to interpret because it is plausible that

https://doi.org/10.1038/s41562-024-01965-x

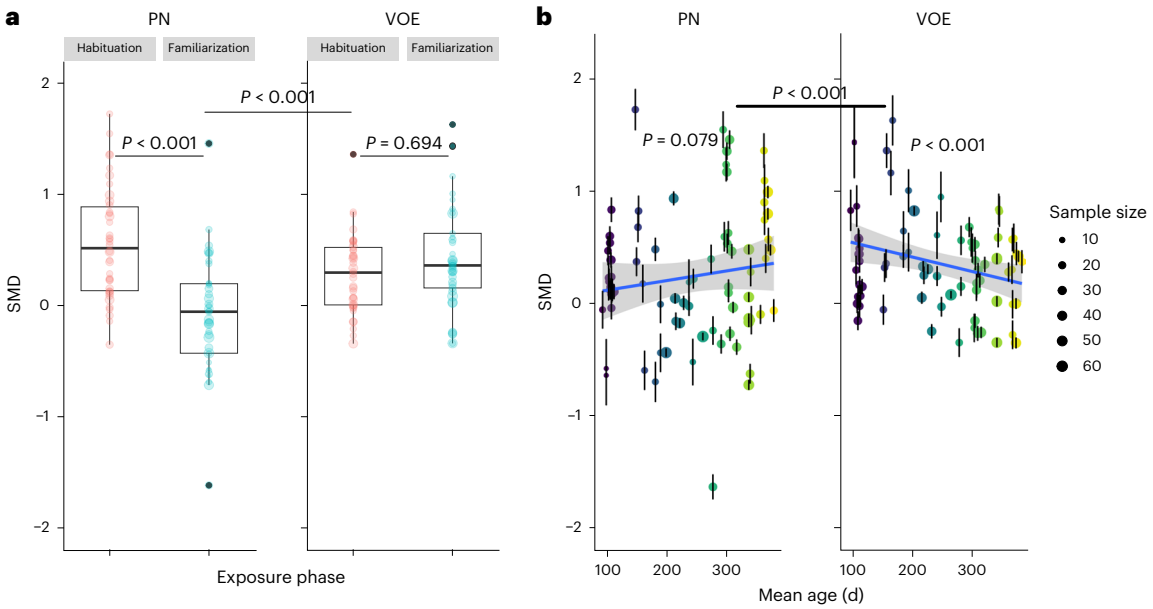


Fig. 6 | PN and VOE are predicted by distinct moderators. a,**b**, Exposure phase (habituation versus familiarization) uniquely predicted the size of the PN effect (**a**) and infant age uniquely predicted the size of the VOE effect (**b**) (for both panels, $n = 1{,}899$ 3- to 12-month-old infants from 76 studies). The data points indicate effect sizes (in SMD) and point sizes indicate sample size. In **a**, the centre of each box indicates the median, the bounds of each box correspond to the 25th and 75th percentiles (IQR) and the whiskers extend to the minima and maxima (up to 1.5× the IQR from the 25th and 75th percentiles). Data beyond the ends of the whiskers are plotted in dark grey. In **b**, the error bars around each data point indicate sampling variance ($\text{s.d.}^2/n$). Blue lines and grey ribbons indicate lines of best fit and 95% CIs, respectively, estimated using a linear model. The quartiles and best-fit lines are unweighted (do not account for differences in sample size or variance across studies). The two-tailed P values were calculated based on mixed effects models including an interaction between trial type and exposure phase (for **a**) and between trial type and infant age (for **b**), along with the other predictors listed in Fig. 5 as additive fixed effects, and a random intercept for each study. We found a significant interaction between exposure phase and trial type (**a**; $X^2(\text{d.f.}) = 134.057(2)$; $P < 0.001$) and between infant age and trial type (**b**; $X^2(\text{d.f.}) = 18.039(2)$; $P < 0.001$). Familiarization studies reported a smaller PN effect to habituation studies (estimate = −4.568 s; $z = -10.136$; $P < 0.001$; 95% CI = (−5.452, −3.685)), with no differences for the VOE effect (estimate = −0.196 s; $z = -0.394$; $P = 0.694$; 95% CI = (−1.169, 0.778)), and infant age was associated with the size of the VOE effect (estimate = −1.323 s; $z = -4.223$; $P < 0.001$; 95% CI = (−1.937, −0.709)) but not the PN effect (estimate = 0.449 s; $z = 1.755$; $P = 0.079$; 95% CI = (−0.053, 0.951)).

researchers designed experiments with more subtle visual differences for older infants. We also found that younger infants tended to show a bigger VOE effect than older infants, but we cannot infer that the VOE effect decreases with age because researchers may design simpler experiments for younger infants. Instead, our results highlight moderators that could be targets of future experimental work. Longitudinal studies of the same infants participating in experiments across time[52,53] or cross-sectional studies testing infants from a large age range using the same stimuli and study design[54] could provide a stronger test of the hypothesis that the VOE effect changes in size over development.

Third, our search criteria, which were applied without reference to the results of the papers, yielded several papers reporting familiarity effects (longer looking towards expected or visually familiar stimuli; for example, refs. 55,56). However, these (and many other) papers did not contain the values required for estimating both the VOE and PN effects—a criterion we pre-registered because we wanted to equate all participant and methodological variables across the two effects. The validity of meta-analyses conducted in our field hinges on the availability of findable, accessible, interoperable and reusable data[57]. Given the slow and labour-intensive nature of infant research, which leads to smaller samples and noisy estimates from each sample, it is all the more important that researchers share anonymized and well-documented data, by default, to enable cumulative science.

Across studies, different moderators explained variance in the size of the PN and VOE effects estimated from the same data. Neuroimaging could be used to more directly test the hypothesis that distinct mechanisms underlie these two looking behaviours. Recent neuroimaging work in adults, using stimuli from this literature[58], showed that early visual regions do not encode prediction error over physical and social expectations. Instead, unexpected stimuli from this literature evoked activity in regions that are associated with domain-specific processing and goal-driven attention[59–61]. The same experiments could be repeated in infants, ideally relating neural correlates of lower- and higher-level prediction error to looking behaviour.

Long before infants can talk or crawl, they explore the world by looking. What are the strengths and challenges of using looking behaviour to study infants' minds? In this Article, we found evidence against the allegations that longer looking towards surprising physical and social events results from an overinterpretation of small noisy samples or is reducible to responses to low-level stimulus features. By aggregating and analysing a large dataset from prior research, we found that infant looking is driven by perceptual and conceptual novelty to a similar degree, and based on distinct predictors. These findings suggest that infant looking behaviour is guided independently by expectation- and stimulus-driven novelty.

Methods

Eligibility for study inclusion

Plans for data curation were pre-registered on the Open Science Framework (OSF; https://osf.io/jghc3) in July 2022. Our goal was to estimate the VOE and PN effects in the same experiments. We also wanted to focus the current research on two of the oldest topics in infant cognition (the understanding of agents and objects) in a sample of studies containing enough variability for us to estimate moderator effects. Thus, we conducted a systematic literature review, specifying the following inclusion criteria: all literature, journal papers, theses, proceedings papers and unpublished datasets after 1985 that tested typically developing infants between 3 and 12 months of age on expectations

about solid objects or single agents engaging in intentional action and that employed an experimental design similar to those shown in Fig. 1. Specifically, to be included: there had to be at least two habituation or familiarization trials before the test trials; the expected and unexpected events had to be equally perceptually novel relative to the previous trials, or the expected event had to be more perceptually novel than the unexpected event; and if the maximum duration of the familiarization and test trials differed, the former had to be fairly long (at least 30 s) or the ratio between maximum durations had to exceed 0.8.

All perceptual differences between the familiarization or habituation and test events, including changes in the path of an agent or object and changes in the location of objects or obstacles (other than those distinguishing expected from unexpected events), could count towards PN. Our main research question was whether these two classes of visual differences (those that make events unexpected and those that make events visually new) are interchangeable or distinct.

Information sources and search strategy

We followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 guidelines[62] when selecting and reporting the studies to be included in our meta-analysis (see Fig. 2 and Extended Data Fig. 7). We started with our own expert knowledge about infant VOE studies in the domains of early physical and psychological reasoning. We also searched for papers including the terms 'violation of expectation', 'habituation or familiarization', 'preferential looking', 'looking time', 'physics or psychology' and 'object or agent' in July to August 2022 on Google Scholar, Proquest and PubMed. We also scanned the reference sections of review articles to locate additional studies and used Google Scholar to identify related articles and articles that cited the first papers testing psychological and physical expectations[5,6,8,63]. Lastly, we emailed two listservs (the Cognitive Development Society and International Congress of Infant Studies) to collect more papers and datasets, including unpublished datasets. We scanned paper titles and abstracts to screen out irrelevant studies, along with the tables, figures, primary text and supplementary materials of each of the potentially relevant papers.

In addition to the criteria from our literature review, for a paper to be included in our final sample it had to report six key values for our confirmatory analysis (the mean and variance of the last habituation or familiarization trial, the first expected trial and the first unexpected trial). For papers that were relevant but otherwise missing one or more pieces of data, we emailed the authors asking them to share the original datasets and included papers for which the original data were found and sent to us. This search process, including the papers screened and authors contacted, is fully documented at https://osf.io/mpkau. This resulted in 33 papers (28 published papers and five unpublished papers) that met our full inclusion criteria, excluding one outlier paper[30], which heavily skewed some of the supplementary meta-analytic results due to its extremely low variance relative to the other studies. Our decision to include or exclude a study in our analysis did not depend on its findings. We did not apply any judgement about whether each included effect was likely to reflect a true or false positive or negative; therefore, our sample of studies may include a mix of all of these.

Data collection process

Each paper that passed through the selection process and contained the required data was coded independently by two team members (L.K. and S.H.P.). Disagreements were resolved by discussion or by consulting with a third team member (S.L.). All team members were aware of the purposes of the study. Papers were screened first by method and then by available data; looking time data were collected last to minimize researcher bias. Analyses were pre-registered after data annotation was complete but before visualizing or analysing the data. When exact numerical values were not provided in a table or the main text but were reported in a figure we used the tool WebPlotDigitizer[64] (https://automeris.io/WebPlotDigitizer/) to extract approximate values from the figures. If there were discrepancies across these data sources (for example, the numbers reported in the paper differed from the numbers calculated using the raw data), the team contacted the authors for resolution. If there were discrepancies between the paper, figures or raw data, we prioritized the data sources in the following order: values from the raw data if available, then author correspondence, then numerical values reported in the paper and finally estimates from figures. Notes about discrepancies in coding and data sources have been fully catalogued and are available on OSF at https://osf.io/4n82q.

Risk assessment

To assess the quality of the data, we coded, for each study, whether the person who generated the looking times per trial was naive at least to the order of test events that infants watched, and whether the paper reported information about inter-rater reliability. Information about each study can be found in Supplementary Table 16.

We found that 74.3% of studies reported that the data were generated by a naive human coder; 23.9% of studies did not specify whether the human coder was naive and the remaining 1.8% were eye-tracking studies that did not use human coding. We note that this percentage is a conservative estimate, because we coded 'yes' for this feature only if the authors explicitly mentioned the naiveness of the human rater. Some papers reported methods that make it likely that coders were naive (for example, they saw a camera feed of the infant's face from another room or looked through a peephole in the puppet stage at the infant), but did not explicitly mention experimenter or observer masking or blinding. We found that 93.4% of studies had a second coder check the reliability of the data and reported information about inter-rater reliability; 4.6% of studies did not report this information and the remaining 1.8% were eye-tracking studies that did not use human coding.

Analysis overview

Plans for data analysis were pre-registered on the OSF (https://osf.io/jghc3/) in March 2023, including several updates. Our goal was to conduct these analyses on as many studies as possible and on both study- and infant-level data. Thus, our pre-registered analysis plan included three steps. First, we estimated the size of the VOE and PN effects in the study-level data (meta-analysis). Second, we examined which participant and experiment features (for example, infant age or study design) predicted the size of each of these two effects. Third, we repeated these same analyses on studies with available data from individual infants (mega-analysis); these data also gave us the opportunity to conduct exploratory individual differences analyses to evaluate the conclusions from the meta-analysis.

Dependent variables. Each included study (for the meta-analysis) or each infant (for the mega-analysis) contributed data for three consecutive trials: the last habituation or familiarization trial, the first expected test event and the first unexpected test event. For all studies in our analysis, the expected test events contained some low-level change in the stimulus. Thus, we defined PN as the difference between the last habituation or familiarization trial and the first expected test event, where a positive value indicated longer looking towards the visually novel event. For the studies in our analyses, testing for a positive VOE effect, the unexpected test events contained a violation of a hypothesized expectation and were either equally or more visually familiar than the expected test event. Thus, we defined VOE as the difference between the first unexpected and first expected test events, where a positive value indicates longer looking towards the unexpected event. For both effects, we chose to measure infants' first exposure to each source of novelty, rather than averaging looks across multiple test trials, to maximize our chances of measuring each effect and to equate the number of trials of each type (the last habituation/familiarization

event only occurred once and test events were often repeated). See Extended Data Fig. 1.

Looking duration in VOE studies is often log-normally distributed (see Extended Data Fig. 8 for the distributions of the PN and VOE effects in our dataset)[33]. In study-level data, we tested the hypothesis that the VOE and PN effects are best expressed as a ratio of means, rather than an SMD. In infant-level data, using the fitdistr() function from the MASS[65] package, we found that a log-normal distribution better fit the distribution of looking times than a normal distribution (log-likelihood for log-normal distribution = −21,562 versus normal distribution = −24,370), so we log-transformed looking durations before our mega-analyses.

Modelling overview. Analyses were conducted in R[66]. All random effects meta-analyses, which explicitly model between-study heterogeneity and assume that effects can truly vary across studies, were carried out using the metafor package[67]. All mega-analyses were carried out using the lme4 package[68] and *P* values for linear mixed effects models were calculated using the lmerTest package[69]. Quality assurance of linear mixed effects models was carried out using the performance[70] package. Analyses of publication bias were carried out using the weightr[71] and metafor[67] packages. Figures 3, 4 and 6 and Extended Data Figs. 1–6 and 8 were made using the ggplot2 (ref. 72), patchwork[73] and viridis[74] packages. Figure 5 was made using the sjPlot[75] package. For all pre-registered analyses, our significance threshold was 0.05 and tests for statistical significance were two tailed. Coefficients and 95% CIs were re-configured so that positive values indicate longer looking towards unexpected than expected events (for the VOE effect) and longer looking towards visually novel than visually familiar events (for the PN effect). Model comparison was carried out on statistical models that were estimated using the maximum likelihood method.

Estimating effect sizes (confirmatory). First, we estimated the size of the VOE and PN effects in study-level data. Average looking time and sampling variance (s.d.^{2}/n) per trial type (last habituation/familiarization, first expected or first unexpected) per study were the outcome variables. Trial type was the only moderator. The first expected test trial was set as the reference level, so that the two estimated model coefficients correspond to the VOE effect (unexpected − expected) and the PN effect (expected − last habituation or familiarization trial). We included one random intercept for study to account for observations nested within studies. Our function call was: rma.mv(yi = mean_looking_time, v = variance_looking_time, mods = ~trial_type, random = ~1|study, data). To evaluate evidence for and against the hypothesis that the VOE and PN effects differ in size, we first computed effect sizes in SMDs for each study using metafor::escalc(). We then fit two mixed effects models with these effect sizes as the dependent measure. The first model included effect type (VOE versus PN) as a fixed effect (representing the hypothesis that these two effects differ in size) and the second model included just the intercept (representing the hypothesis that these two effects are similar in size). Both included a random intercept for study ID. We then used Wagenmaker's method to compute Bayes factors between these two models based on their respective BICs[76].

Publication bias (exploratory). We conducted two exploratory and complementary tests for publication bias, which could take the form of selective reporting, the file drawer problem and/or journals unwilling to publish null or unclear findings. First, we conducted Egger's test for funnel plot asymmetry, which tests for the presence of low-precision studies with more positive or large effects than more negative or small effects[77]. Then, we used selection models to test whether there was an over-representation of significant results[78] using weightr::weightfunct(). For a discussion of the strengths and weaknesses of both methods, see ref. 32.

Form of looking preferences (exploratory). Next, we asked whether the VOE and PN effects are additive or multiplicative. Using metafor::escalc(), we computed effect sizes for each effect for each study in two different ways: as an SMD and as a log ratio of means. Then, each of these effect sizes was modelled as the intercept in a random effects meta-analysis with no moderators (formula: rma.mv(yi, vi, random = ~1|study, data)). We then used metafor::fitstats() to generate the AIC and BIC values of both models and compared these values across models to assess whether each effect was better expressed as an SMD or ratio of means. We converted model coefficients and values in CIs from the log ratio of means to the ratio of means for ease of interpretability.

Analyses of moderators (pre-registered exploratory). We then asked whether similar or different predictors moderated the size of the PN and VOE effects. We first fit separate meta-analytic mixed effects models for each effect. The model specified the SMD (yi) and sampling variance (vi) as the dependent variables and the fixed effects listed in Table 2 as predictors, including an additional random intercept for study. We *z* scored age across studies before entering it as a predictor and used summed contrasts for all categorical predictors so that the model intercept could be interpreted as the grand mean of the PN or VOE effect (in SMD units) at the mean age of the infants. The function call for the moderator analyses was: rma.mv(yi, vi, mods = ~equal_per_nov + exposure_phase + scale(mean_age) + domain + stim_loop, random = ~1|study, data).

To study whether the significant moderators of the VOE and PN effects, originally modelled separately, differentially predicted each effect, we estimated both effects in the same model with: average looking and associated sampling variance as the outcome variables; an interaction between (1) trial type and (2) the moderator in question as the key fixed effect; and the other moderators as additive fixed effects. For example, a function call assessing whether exposure phase differentially predicted the PN and VOE effects was: rma.mv(yi, vi, mods = ~trial_type * exposure_phase + equal_per_nov + scale(mean_age) + domain + stim_loop, random = ~1|study, data).

Mega-analytic analogues (pre-registered exploratory). We also fit linear mixed effects models over data from individual infants (25 papers, 60 studies and 1,482 infants). The dependent variable for all models was the looking duration per trial per infant in log seconds. The fixed effects were identical to those described above, except that these models included ages of individual infants *z*-scored across studies. We also added test trial order as an additional between-participants fixed effect. All models that estimated the VOE and PN effects separately used a difference score per infant per effect as the dependent variable (for example, for PN, the difference between looking at the first expected and last habituation trials, in log seconds). All models that estimated the VOE and PN effects together included looking, in log seconds, for each trial type per infant as the dependent variable. Separate models of the PN and VOE effects, to which each infant contributed one difference score, included random intercepts to account for observations within studies and experiments ((1|study_ID) + (1|expt_ID)). (Experiments often randomly assigned infants to separate sub-experiments; study_ID refers to these sub-experiments and expt_ID refers to the broader category.) Models estimating the size of both effects simultaneously, to which each infant contributed three looking times, included an additional intercept for participant ID, to account for repeated measures within infants ((1|specific_subject_ID) + (1|study_ID) + (1|expt_ID)). These analyses and their results are described in full in the Supplementary Information.

Degrees of freedom were estimated using the Satterthwaite approximation method. Standardized betas (β) express the size of each effect in standard deviation and were computed by *z* scoring all continuous variables in the model. In further exploratory analyses, we studied the size of the PN effect in negative control experiments.

We also conducted a power analysis by estimating the number of infants required in a new sample to measure a novel VOE and PN effect. These analyses and their results are described in full in the Supplementary Information.

Individual differences in habituation rate and VOE and PN effects (exploratory). We studied the relationship between habituation, VOE and PN using data from individual infants in habituation studies (22 studies; $n = 499$ infants) and familiarization studies with at least six familiarization trials (21 studies; $n = 603$ infants). We fit separate linear mixed effects models for habituation and familiarization studies, with looking time in log seconds from individual infants during the last habituation, first expected and first unexpected trials as the dependent measure. Both models included random intercepts to account for differences in looking behaviour across participants, studies and experiments. For the habituation studies, our model included an interaction between trial type and the number of trials infants underwent before reaching the habituation criteria set by the original authors (model formula: looking_time ~ trial_type × num_hab_trials + (1|specific_subject_ID) + (1|study_ID) + (1|expt_ID)). For the familiarization studies, our model included an interaction between trial type and a binary predictor that indicated whether each infant would have met a standard habituation criterion (total looking on the last three familiarization trials that was 50% or less than the total looking on the first three familiarization trials, in seconds, by the end of familiarization (model formula: looking_time ~ trial_type × habituated + (1|specific_subject_ID) + (1|study_ID) + (1|expt_ID))). The full results of these analyses can be found in the Supplementary Information.

Quality assurance. In analyses of moderators in infant-level data, we tested for collinearity between all of the fixed effects using check_collinearity() from the performance package[70] and found no evidence for collinearity issues (for all main effects, variance inflation factors ranged from ~1–2 and tolerance from ~0.5–1.0). Thus, we proceeded as planned and included all of them in our model. After fitting each model, we used check_models() from the performance package[70] to check the normality of residuals and random effects, the homogeneity of variance and influential observations of all mega-analytic models. The outputs of these checks can be accessed at https://osf.io/b59km/. For models with influential observations, removing these observations did not change the interpretation of the results.

Deviations from pre-registration. We originally preregistered that we would include paper (1|paper_ID) as a random intercept in the mega-analysis models. Including this random effect led to a failure to converge in one of our models, so we removed it from all mega-analysis models to maintain consistency. We also excluded one paper[30] from the results in the main text, the inclusion of which led to issues in supplementary meta-analyses due to extremely low variance. The results including this paper are presented in the Supplementary Information for full transparency. All of our primary results hold, regardless of whether this paper is excluded or included.

This project grew from a small-scale analysis of one dataset[12] ($n = 60$ infants). During that initial case study, we pre-registered that we would define the VOE effect as the difference between the unexpected and last habituation event. Before collecting or analysing the remaining 97% of the data, we pre-registered equations (1) and (2) as they appear in the main text.

Ethics approval. This meta-analysis reports data from published and unpublished previous research. Each of these datasets was collected with approval from the institutional review board of the corresponding university; original data provided by the authors of this research do not include any personally identifiable information. The current research did not involve any new data collection or interaction with or intervention involving human participants and therefore does not meet the definition of human participants research according to the US Department of Health and Human Services (https://www.ecfr.gov/on/2018-07-19/title-45/part-46#p-46.102(e)).

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All anonymized data associated with this paper are openly available at https://osf.io/b59km/ and from Zenodo (https://doi.org/10.5281/zenodo.12629030)[79].

Code availability

All analysis scripts associated with this paper are openly available at https://osf.io/b59km/ and from Zenodo (https://doi.org/10.5281/zenodo.12629030)[79].

References

1. Baillargeon, R., Scott, R. M. & Bian, L. Psychological reasoning in infancy. *Annu. Rev. Psychol.* **67**, 159–186 (2016).
2. Hespos, S. J. & vanMarle, K. Physics for infants: characterizing the origins of knowledge about objects, substances, and number. *Wiley Interdiscip. Rev. Cogn. Sci.* **3**, 19–27 (2012).
3. Spelke, E. S. *What Babies Know: Core Knowledge and Composition* Vol. 1 (Oxford Univ. Press, 2022).
4. Margoni, F., Surian, L. & Baillargeon, R. The violation-of-expectation paradigm: a conceptual overview. *Psychol. Rev.* **131**, 716–748 (2024).
5. Baillargeon, R., Spelke, E. S. & Wasserman, S. Object permanence in five-month-old infants. *Cognition* **20**, 191–208 (1985).
6. Gergely, G., Nádasdy, Z., Csibra, G. & Bíró, S. Taking the intentional stance at 12 months of age. *Cognition* **56**, 165–193 (1995).
7. Needham, A. & Baillargeon, R. Intuitions about support in 4.5-month-old infants. *Cognition* **47**, 121–148 (1993).
8. Woodward, A. L. Infants selectively encode the goal object of an actor's reach. *Cognition* **69**, 1–34 (1998).
9. Blumberg, M. S. & Adolph, K. E. Protracted development of motor cortex constrains rich interpretations of infant cognition. *Trends Cogn. Sci.* **27**, 233–245 (2023).
10. Cohen, L. B. & Marks, K. S. How infants process addition and subtraction events. *Dev. Sci.* **5**, 186–201 (2002).
11. Paulus, M. Should infant psychology rely on the violation-of-expectation method? Not anymore. *Infant Child Dev.* **31**, e2306 (2022).
12. Liu, S. & Spelke, E. S. Six-month-old infants expect agents to minimize the cost of their actions. *Cognition* **160**, 35–42 (2017).
13. Byers-Heinlein, K., Bergmann, C. & Savalei, V. Six solutions for more reliable infant research. *Infant Child Dev.* **31**, e2296 (2022).
14. Cohen, L. B. Uses and misuses of habituation and related preference paradigms. *Infant Child Dev.* **13**, 349–352 (2004).
15. Haith, M. M. Who put the cog in infant cognition? Is rich interpretation too costly? *Infant Behav. Dev.* **21**, 167–179 (1998).
16. Jackson, I. R. & Sirois, S. But that's possible! Infants, pupils, and impossible events. *Infant Behav. Dev.* **67**, 101710 (2022).
17. Bornstein, M. H., Kessen, W. & Weiskopf, S. Color vision and hue categorization in young human infants. *J. Exp. Psychol. Hum. Percept. Perform.* **2**, 115–129 (1976).
18. Caron, A. J., Caron, R. F. & Carlson, V. R. Infant perception of the invariant shape of objects varying in slant. *Child Dev.* **50**, 716–721 (1979).
19. Kessen, W. & Bornstein, M. H. Discriminability of brightness change for infants. *J. Exp. Child Psychol.* **25**, 526–530 (1978).

20. Braddick, O. J., Atkinson, J. & Wattam-Bell, J. R. Development of the discrimination of spatial phase in infancy. *Vis. Res.* **26**, 1223–1239 (1986).
21. Wattam-Bell, J. Visual motion processing in one-month-old infants: habituation experiments. *Vis. Res.* **36**, 1679–1685 (1996).
22. Bergmann, C. & Cristia, A. Development of infants' segmentation of words from native speech: a meta-analytic approach. *Dev. Sci.* **19**, 901–917 (2016).
23. Enge, A., Kapoor, S., Kieslinger, A.-S. & Skeide, M. A. A meta-analysis of mental rotation in the first years of life. *Dev. Sci.* **26**, e13381 (2023).
24. Rabagliati, H., Ferguson, B. & Lew-Williams, C. The profile of abstract rule learning in infancy: meta-analytic and experimental evidence. *Dev. Sci.* **22**, e12704 (2019).
25. Bergmann, C., Rabagliati, H. & Tsuji, S. What's in a looking time preference? Preprint at *PsyArXiv* https://doi.org/10.31234/osf.io/6u453 (2019).
26. Kosie, J. E. et al. ManyBabies 5: a large-scale investigation of the proposed shift from familiarity preference to novelty preference in infant looking time. Preprint at *PsyArXiv* https://doi.org/10.31234/osf.io/ck3vd (2023).
27. Koile, E. & Cristia, A. Toward cumulative cognitive science: a comparison of meta-analysis, mega-analysis, and hybrid approaches. *Open Mind* **5**, 154–173 (2021).
28. Bergmann, C. et al. Promoting replicability in developmental research through meta-analyses: insights from language acquisition research. *Child Dev.* **89**, 1996–2009 (2018).
29. Singh, L., Cristia, A., Karasik, L. B., Rajendra, S. J. & Oakes, L. M. Diversity and representation in infant research: barriers and bridges toward a globalized science of infant development. *Infancy* **28**, 708–737 (2023).
30. Sanal-Hayes, N. E. M., Hayes, L. D., Walker, P., Mair, J. L. & Bremner, J. G. Adults' understanding and 6-to-7-month-old infants' perception of size and mass relationships in collision events. *Appl. Sci.* **12**, 9846 (2022).
31. Duval, S. & Tweedie, R. A nonparametric 'trim and fill' method of accounting for publication bias in meta-analysis. *J. Am. Stat. Assoc.* **95**, 89–98 (2000).
32. Maier, M., VanderWeele, T. J. & Mathur, M. B. Using selection models to assess sensitivity to publication bias: a tutorial and call for more routine use. *Campbell Syst. Rev.* **18**, e1256 (2022).
33. Csibra, G., Hernik, M., Mascaro, O., Tatone, D. & Lengyel, M. Statistical treatment of looking-time data. *Dev. Psychol.* **52**, 521–536 (2016).
34. Richards, J. E. & Gibson, T. L. Extended visual fixation in young infants: look distributions, heart rate changes, and attention. *Child Dev.* **68**, 1041–1056 (1997).
35. Šimkovic, M. & Träuble, B. Additive and multiplicative probabilistic models of infant looking times. *PeerJ* **9**, e11771 (2021).
36. Van den Noortgate, W., López-López, J. A., Marín-Martínez, F. & Sánchez-Meca, J. Three-level meta-analysis of dependent effect sizes. *Behav. Res. Methods* **45**, 576–594 (2012).
37. Oakes, L. M. Using habituation of looking time to assess mental processes in infancy. *J. Cogn. Dev.* **11**, 255–268 (2010).
38. Aslin, R. N. What's in a look? *Dev. Sci.* **10**, 48–53 (2007).
39. Stahl, A. E. & Kibbe, M. M. Great expectations: the construct validity of the violation-of-expectation method for studying infant cognition. *Infant Child Dev.* **31**, e2359 (2022).
40. Sirois, S. & Mareschal, D. Models of habituation in infancy. *Trends Cogn. Sci.* **6**, 293–298 (2002).
41. Boisseau, R. P., Vogel, D. & Dussutour, A. Habituation in non-neural organisms: evidence from slime moulds. *Proc. R. Soc. B Biol. Sci.* **283**, 20160446 (2016).
42. Turatto, M., Bonetti, F., Chiandetti, C. & Pascucci, D. Context-specific distractors rejection: contextual cues control long-term habituation of attentional capture by abrupt onsets. *Vis. Cogn.* **27**, 291–304 (2019).
43. Colombo, J. & Mitchell, D. W. Infant visual habituation. *Neurobiol. Learn. Mem.* **92**, 225–234 (2009).
44. Thompson, R. F. Habituation: a history. *Neurobiol. Learn. Mem.* **92**, 127–134 (2009).
45. Saffran, J. R. & Kirkham, N. Z. Infant statistical learning. *Annu. Rev. Psychol.* **69**, 181–203 (2018).
46. Stojnić, G., Gandhi, K., Yasuda, S., Lake, B. M. & Dillon, M. R. Commonsense psychology in human infants and machines. *Cognition* **235**, 105406 (2023).
47. Shu, T. et al. AGENT: a benchmark for core psychological reasoning. In *Proc. 38th International Conference on Machine Learning* 9614–9625 (PMLR, 2021).
48. Liu, S., Ullman, T. D., Tenenbaum, J. B. & Spelke, E. S. Ten-month-old infants infer the value of goals from the costs of actions. *Science* **358**, 1038–1041 (2017).
49. Kidd, C., Piantadosi, S. T. & Aslin, R. N. The goldilocks effect: human infants allocate attention to visual sequences that are neither too simple nor too complex. *PLoS ONE* **7**, e36399 (2012).
50. Piloto, L. S., Weinstein, A., Battaglia, P. & Botvinick, M. Intuitive physics learning in a deep-learning model inspired by developmental psychology. *Nat. Hum. Behav.* **6**, 1257–1267 (2022).
51. Smith, K. A. et al. Modeling expectation violation in intuitive physics with coarse probabilistic object representations. In *Proc. 33rd Conference on Neural Information Processing Systems* (eds Wallach, H. et al.) 8983–8993 (Curran Associates, 2019).
52. Adolph, K. E., Robinson, S. R., Young, J. W. & Gill-Alvarez, F. What is the shape of developmental change? *Psychol. Rev.* **115**, 527–543 (2008).
53. Siegler, R. S. & Crowley, K. The microgenetic method: a direct means for studying cognitive development. *Am. Psychol.* **46**, 606–620 (1991).
54. Frank, M. C., Braginsky, M., Yurovsky, D. & Marchman, V. A. Wordbank: an open repository for developmental vocabulary data. *J. Child Lang.* **44**, 677–694 (2017).
55. Bogartz, R. S., Shinskey, J. L. & Schilling, T. H. Object permanence in five-and-a-half-month-old infants? *Infancy* **1**, 403–428 (2000).
56. Schilling, T. H. Infants' looking at possible and impossible screen rotations: the role of familiarization. *Infancy* **1**, 389–402 (2000).
57. Wilkinson, M. D. et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **3**, 160018 (2016).
58. Liu, S., Lydic, K., Mei, L. & Saxe, R. Violations of physical and psychological expectations in the human adult brain. *Imaging Neurosci.* **2**, 1–25 (2024).
59. Pramod, R., Cohen, M. A., Tenenbaum, J. B. & Kanwisher, N. Invariant representation of physical stability in the human brain. *eLife* **11**, e71736 (2022).
60. Fischer, J., Mikhael, J. G., Tenenbaum, J. B. & Kanwisher, N. Functional neuroanatomy of intuitive physical inference. *Proc. Natl Acad. Sci. USA* **113**, E5072–E5081 (2016).
61. Fedorenko, E., Duncan, J. & Kanwisher, N. Broad domain generality in focal regions of frontal and parietal cortex. *Proc. Natl Acad. Sci. USA* **110**, 16616–16621 (2013).
62. Page, M. J. et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *Int. J. Surg.* **88**, 105906 (2021).
63. Spelke, E. S., Breinlinger, K., Macomber, J. & Jacobson, K. Origins of knowledge. *Psychol. Rev.* **99**, 605–632 (1992).
64. Rohatgi, A. WebPlotDigitizer: extract data from plots, images, and maps (version 4.5). *automeris.io* https://automeris.io/WebPlotDigitizer/ (2022).

65. Venables, W. N. & Ripley, B. D. *Modern Applied Statistics with S* (Springer, 2003).
66. R Core Development Team *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2023).
67. Viechtbauer, W. Conducting meta-analyses in R with the metafor package. *J. Stat. Softw.* **36**, 1–48 (2010).
68. Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* **67**, 1–48 (2015).
69. Kuznetsova, A., Brockhoff, P. B. & Christensen, R. H. B. lmerTest package: tests in linear mixed effects models. *J. Stat. Softw.* **82**, 1–26 (2017).
70. Lüdecke, D., Ben-Shachar, M. S., Patil, I., Waggoner, P. & Makowski, D. performance: an R package for assessment, comparison and testing of statistical models. *J. Open Source Softw.* **6**, 3139 (2021).
71. Coburn, K. M. & Vevea, J. L. weightr: estimating weight-function models for publication bias (version 2.0.2). *CRAN* https://CRAN.R-project.org/package=weightr (2016).
72. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis* (Springer-Verlag, 2016).
73. Pedersen, T. L. patchwork: the composer of plots (version 1.1.3). *CRAN* https://CRAN.R-project.org/package=patchwork (2019).
74. Garnier, S. viridis: colorblind-friendly color maps for R (version 0.6.4). *CRAN* https://CRAN.R-project.org/package=viridis (2015).
75. Lüdecke, D. sjPlot: data visualization for statistics in social science (version 2.8.15). *CRAN* https://CRAN.R-project.org/package=sjPlot (2013).
76. Wagenmakers, E.-J. A practical solution to the pervasive problems of *P* values. *Psychon. Bull. Rev.* **14**, 779–804 (2007).
77. Egger, M., Smith, G. D., Schneider, M. & Minder, C. Bias in meta-analysis detected by a simple, graphical test. *Br. Med. J.* **315**, 629–634 (1997).
78. Vevea, J. L. & Hedges, L. V. A general linear model for estimating effect size in the presence of publication bias. *Psychometrika* **60**, 419–435 (1995).
79. Kunin, L., Liu, S., Piccolo, S. & Saxe, R. A systematic meta-analysis of the role of perceptual and conceptual novelty in guiding infant looking behaviour. *Zenodo* https://doi.org/10.5281/zenodo.12629030 (2024).
80. Biro, S., Csibra, G. & Gergely, G. The role of behavioral cues in understanding goal-directed actions in infancy. *Prog. Brain Res.* **164**, 303–322 (2007).
81. Brandone, A. C. & Wellman, H. M. You can't always get what you want: infants understand failed goal-directed actions. *Psychol. Sci.* **20**, 85–91 (2009).
82. Choi, Y., Mou, Y. & Luo, Y. How do 3-month-old infants attribute preferences to a human agent? *J. Exp. Child Psychol.* **172**, 96–106 (2018).
83. Chuey, A. et al. Moderated online data-collection for developmental research: methods and replications. *Front. Psychol.* **12**, 734398 (2021).
84. Gerson, S. A. & Woodward, A. L. The joint role of trained, untrained, and observed actions at the origins of goal recognition. *Infant Behav. Dev.* **37**, 94–104 (2014).
85. Gerson, S. A. & Woodward, A. L. Learning from their own actions: the unique effect of producing actions on infants' action understanding. *Child Dev.* **85**, 264–277 (2014).
86. Hernik, M. & Southgate, V. Nine-months-old infants do not need to know what the agent prefers in order to reason about its goals: on the role of preference and persistence in infants' goal-attribution. *Dev. Sci.* **15**, 714–722 (2012).
87. Hespos, S. J., Ferry, A. L. & Rips, L. J. Five-month-old infants have different expectations for solids and liquids. *Psychol. Sci.* **20**, 603–611 (2009).
88. Lakusta, L. & Carey, S. Twelve-month-old infants' encoding of goal and source paths in agentive and non-agentive motion events. *Lang. Learn. Dev.* **11**, 152–175 (2015).
89. Liu, S., Brooks, N. B. & Spelke, E. S. Origins of the concepts cause, cost, and goal in prereaching infants. *Proc. Natl Acad. Sci. USA* **116**, 17747–17752 (2019).
90. Liu, S. et al. Dangerous ground: one-year-old infants are sensitive to peril in other agents' action plans. *Open Mind* **6**, 211–231 (2022).
91. Liu, S., Piccolo, S. & Saxe, R. Infants' expectations about object solidity and support, and agents' goal-directed actions: online replications. Preprint at *OSF* https://doi.org/10.17605/OSF.IO/JVQDG (2024).
92. Liu, S. & Spelke, E. S. Infants' understanding of inclined planes: replication of Kim & Spelke (1992) using eyetracking. Preprint at *OSF* https://doi.org/10.17605/OSF.IO/TBJ37 (2024).
93. Liu, S. & Spelke, E. S. Infants' expectations about action efficiency and inclined planes: online replications. Preprint at *OSF* https://doi.org/10.17605/OSF.IO/T23X4 (2024).
94. Luo, Y. & Baillargeon, R. Can a self-propelled box have a goal? Psychological reasoning in 5-month-old infants. *Psychol. Sci.* **16**, 601–608 (2005).
95. Luo, Y. & Baillargeon, R. When the ordinary seems unexpected: evidence for incremental physical knowledge in young infants. *Cognition* **95**, 297–328 (2005).
96. Luo, Y. & Johnson, S. C. Recognizing the role of perception in action at 6 months. *Dev. Sci.* **12**, 142–149 (2009).
97. Luo, Y., Kaufman, L. & Baillargeon, R. Young infants' reasoning about physical events involving inert and self-propelled objects. *Cogn. Psychol.* **58**, 441–486 (2009).
98. Luo, Y. Do 8-month-old infants consider situational constraints when interpreting others' gaze as goal-directed action? *Infancy* **15**, 392–419 (2010).
99. Luo, Y. Three-month-old infants attribute goals to a non-human agent. *Dev. Sci.* **14**, 453–460 (2011).
100. Martin, A., Shelton, C. C. & Sommerville, J. A. Once a frog-lover, always a frog-lover?: infants' goal generalization is influenced by the nature of accompanying speech. *J. Exp. Psychol. Gen.* **146**, 859–871 (2017).
101. Olofson, E. L. & Baldwin, D. Infants recognize similar goals across dissimilar actions involving object manipulation. *Cognition* **118**, 258–264 (2011).
102. Schlottmann, A., Ray, E. D. & Surian, L. Emerging perception of causality in action-and-reaction sequences from 4 to 6 months of age: is it domain-specific? *J. Exp. Child Psychol.* **112**, 208–230 (2012).
103. Skerry, A. E., Carey, S. E. & Spelke, E. S. First-person action experience reveals sensitivity to action efficiency in prereaching infants. *Proc. Natl Acad. Sci. USA* **110**, 18728–18733 (2013).
104. Spaepen, E. & Spelke, E. Will any doll do? 12-month-olds' reasoning about goal objects. *Cogn. Psychol.* **54**, 133–154 (2007).
105. Thoermer, C., Woodward, A., Sodian, B., Perst, H. & Kristen, S. To get the grasp: seven-month-olds encode and selectively reproduce goal-directed grasping. *J. Exp. Child Psychol.* **116**, 499–509 (2013).
106. Woo, B. M., Liu, S. & Spelke, E. S. Infants rationally infer the goals of other people's reaches in the absence of first-person experience with reaching actions. *Dev. Sci.* **27**, e13453 (2024).

Acknowledgements

Portions of the current work were included in L.K.'s 2023 MEng thesis at the Massachusetts Institute of Technology. We thank the researchers who contributed data to this project; J. Cetron at the Harvard Institute for Quantitative Social Science and M. Zettersten

for statistical consultation; members of R.S.'s laboratory for helpful discussion; and the MetaLab team (https://langcog.github.io/metalab/) for their toolkit and tutorials on conducting meta-analyses on developmental data; and M. Frank and G. Raz for feedback on an earlier draft of the manuscript. We gratefully acknowledge the following funding sources: the National Institutes of Health (F32HD103363 to S.L.); Defense Advanced Research Projects Agency (CW3013552 to S.H.P. and L.K.); and Massachusetts Institute of Technology Undergraduate Research Opportunities Program (to L.K.). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

L.K. and S.L. planned the research in consultation with S.H.P. and R.S. L.K., S.H.P. and S.L. carried out the research. L.K. analysed the data in consultation with S.L. L.K. and S.L. wrote the original draft of the paper. R.S. and S.H.P. provided critical feedback. S.L. and L.K. revised the paper in consultation with R.S.

Competing interests

The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41562-024-01965-x.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41562-024-01965-x.

Correspondence and requests for materials should be addressed to Shari Liu.

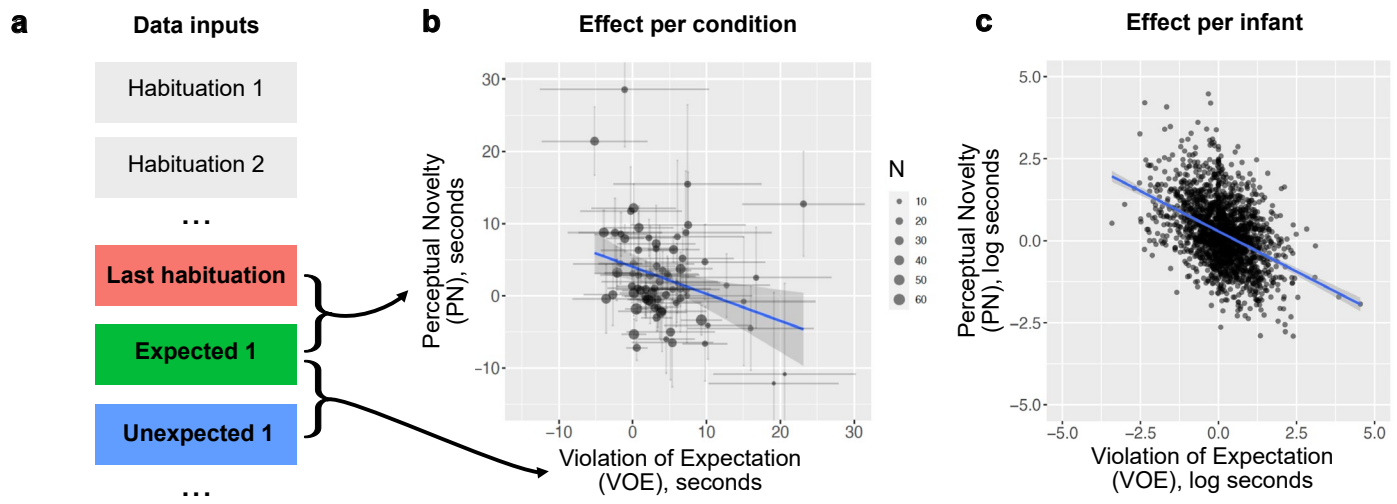
Peer review information *Nature Human Behaviour* thanks Francesco Margoni, Luca Surian and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

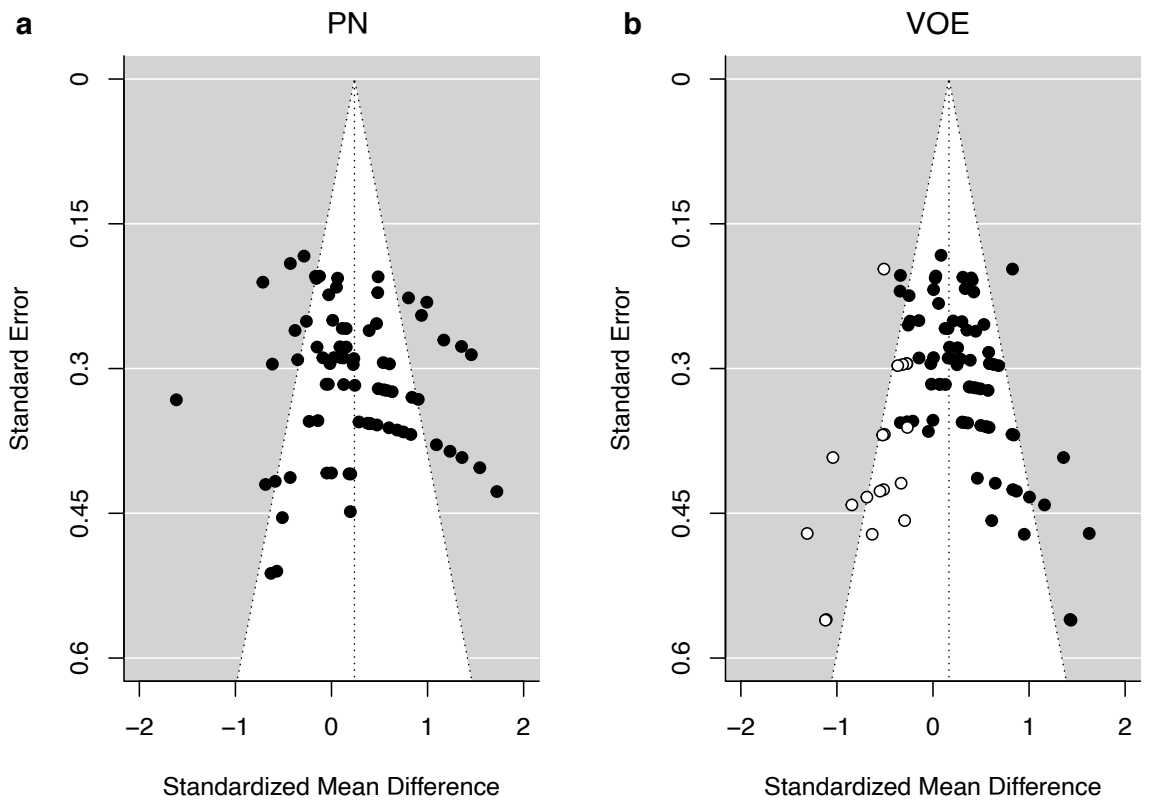
[1]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA. [2]Department of Psychology, Northeastern University, Boston, MA, USA. [3]Department of Psychological and Brain Sciences, Johns Hopkins University, Baltimore, MD, USA. ✉e-mail: sliu199@jhu.edu

https://doi.org/10.1038/s41562-024-01965-x



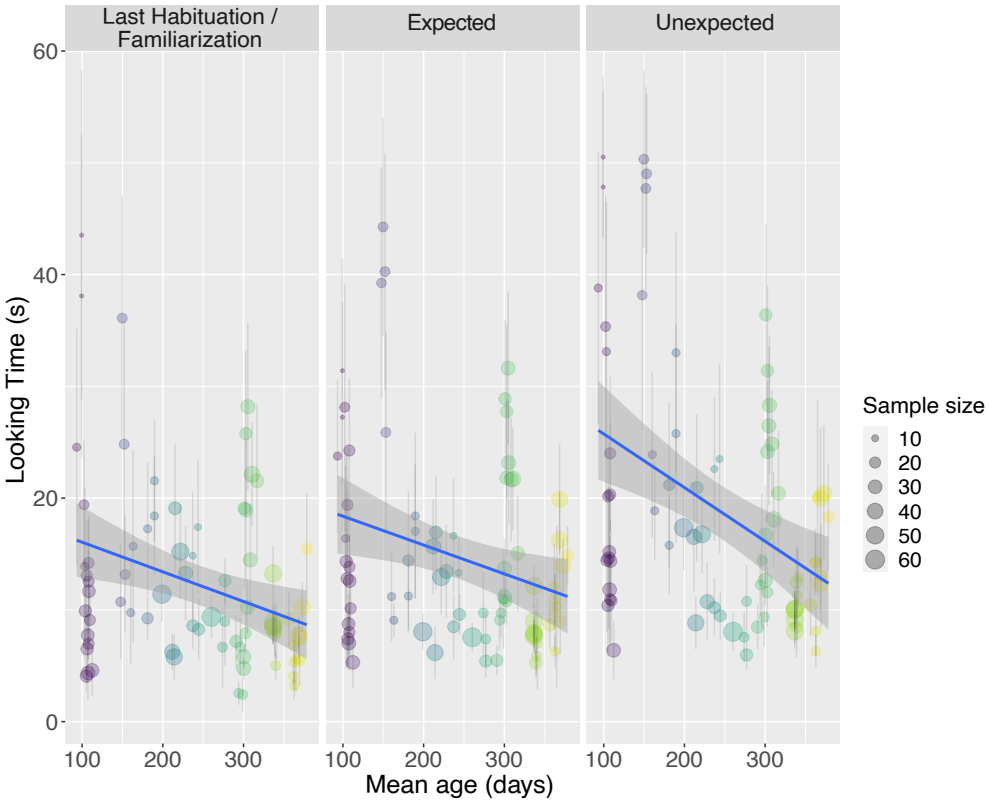
Extended Data Fig. 1 | Procedure for estimating the PN and VOE effects. (**a**) The data inputs, with the three critical trials are in bold (some experiments used habituation, listed here; others used familiarization; and all studies counterbalanced the order of the test events across infants). (**b, c**) The VOE and PN effects plotted against each other, (b) per study (N = 76 studies) or (c) per infant (N = 1482 infants). In (b), error bars around points indicate standard error of the mean (SE), and point size indicates sample size. In (b-c), a best fit line in blue was estimated using a linear model per trial type, and the grey ribbon around the line indicates the 95% confidence intervals.

https://doi.org/10.1038/s41562-024-01965-x



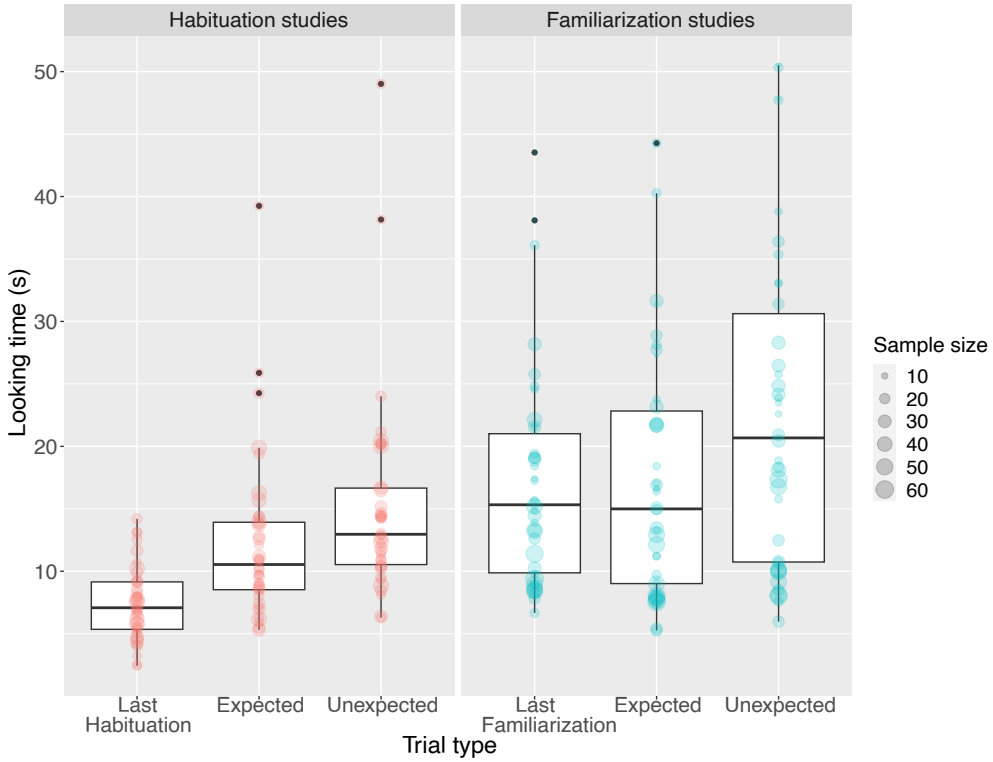
Extended Data Fig. 2 | Funnel plots for PN and VOE effects. Each plot shows effect sizes (standardised mean differences) plotted against the precision (standard error) of each study (N = 76 studies total), for (**a**) the perceptual novelty effect, and (**b**) the violation-of-expectation effect. Black points indicate studies included in our primary analyses; white points were added by the trim-and-fill method to account for possible publication bias. See Methods for details.

https://doi.org/10.1038/s41562-024-01965-x



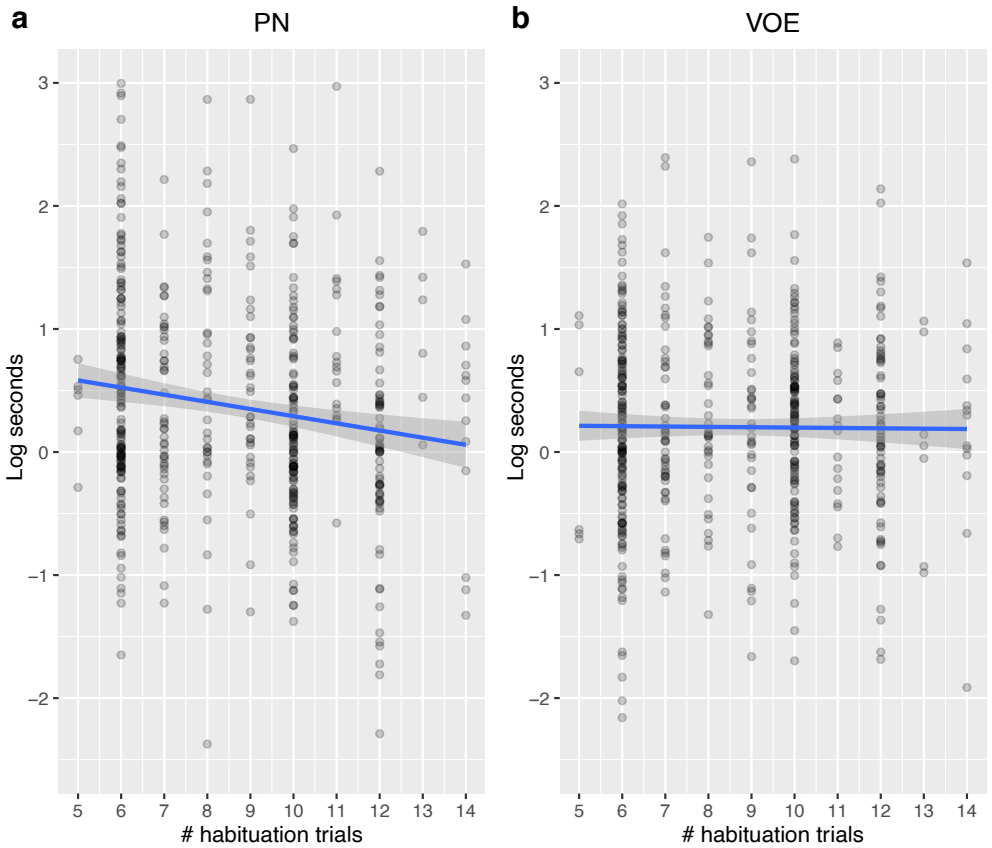
Extended Data Fig. 3 | The relationship between infant age and looking time for each trial type. Each point represents the mean looking time to one trial type for one study (N = 76 studies; 'Last habituation / familiarization' indicates looking on the last trial before test; 'Expected' indicates looking on the first expected test trial; 'Unexpected' indicates looking on the first unexpected test trial.) Point sizes indicate sample sizes. A best fit line estimated using a linear model per trial type is shown in blue, and error bars around points and the grey ribbon around the line indicate 95% confidence intervals. These best fit lines are unweighted (do not take into account differences in the sample sizes or variances across studies).

https://doi.org/10.1038/s41562-024-01965-x



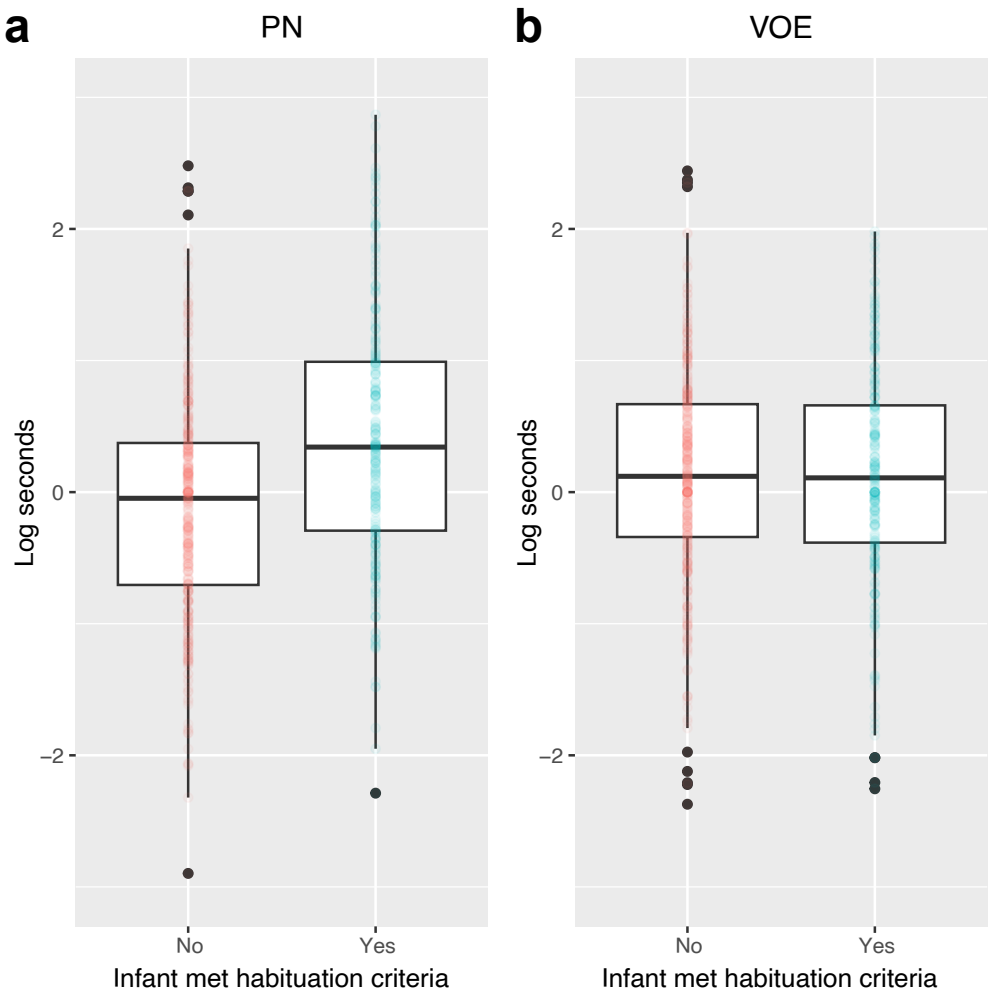
Extended Data Fig. 4 | The relationship between exposure phase and looking time for each trial type. Each point represents the mean looking time to one trial type for one study (N = 76 studies; 'Last habituation / familiarization' indicates looking on the last trial before test; 'Expected' indicates looking on the first expected test trial; 'Unexpected' indicates looking on the first unexpected test trial.). Point sizes indicate sample sizes. The centre of the box indicates the median, the bounds of the box correspond to the 25th and 75th percentiles (the interquartile range, or IQR), and the whiskers extend to the minima and maxima (up to 1.5 IQRs from the 25th and 75th percentiles). Data beyond the end of the whiskers are plotted in dark grey. Quartiles are unweighted (do not account for differences in sample size or variance across studies).

https://doi.org/10.1038/s41562-024-01965-x



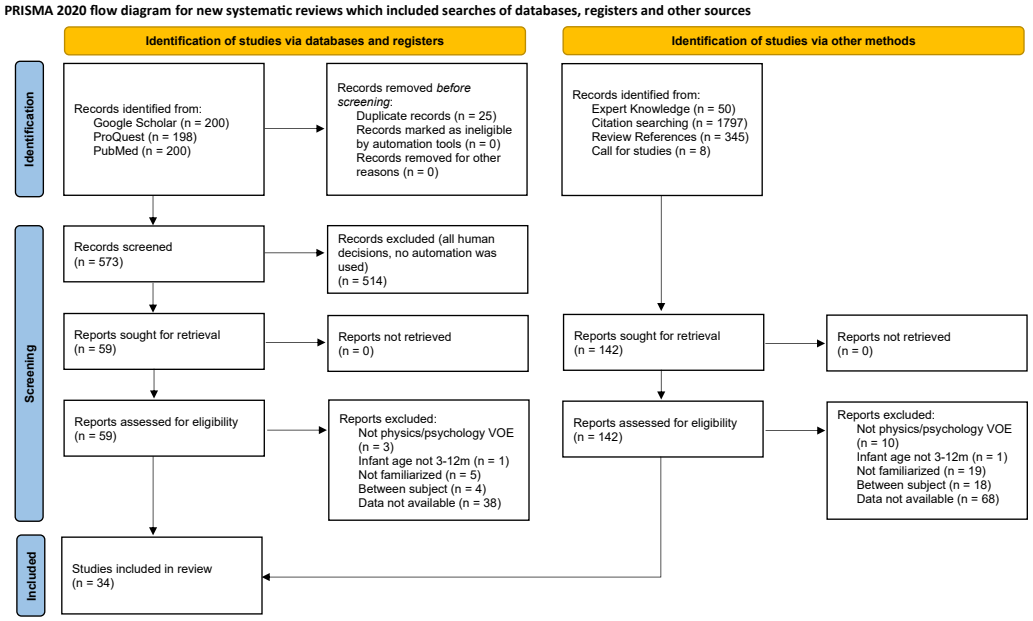
Extended Data Fig. 5 | Relationship between habituation rate and the PN and VOE in individual infants from habituation studies. (**a**) and (**b**) show scatterplots of the (a) PN and (b) VOE effects in log seconds against the number of habituation trials infants saw prior to test trials (22 studies, N = 499 infants). Each point represents one infant's PN and VOE effects. A best fit line estimated using a linear model per trial type is shown in blue, and the grey ribbon around the line indicates 95% confidence intervals.

https://doi.org/10.1038/s41562-024-01965-x



Extended Data Fig. 6 | Relationship between habituation criteria and the PN and VOE in individual infants from familiarization studies. (**a**) and (**b**) show boxplots of the (a) PN and (b) VOE effects in log seconds, broken down by whether infants met a standard habituation criteria by the end of the familiarization phase (21 studies, N = 603 infants). The centre of the box indicates the median, the bounds of the box correspond to the 25th and 75th percentiles (the interquartile range, or IQR), and the whiskers extend to the minima and maxima (up to 1.5 IQRs from the 25th and 75th percentiles). Data beyond the end of the whiskers are plotted in dark grey.

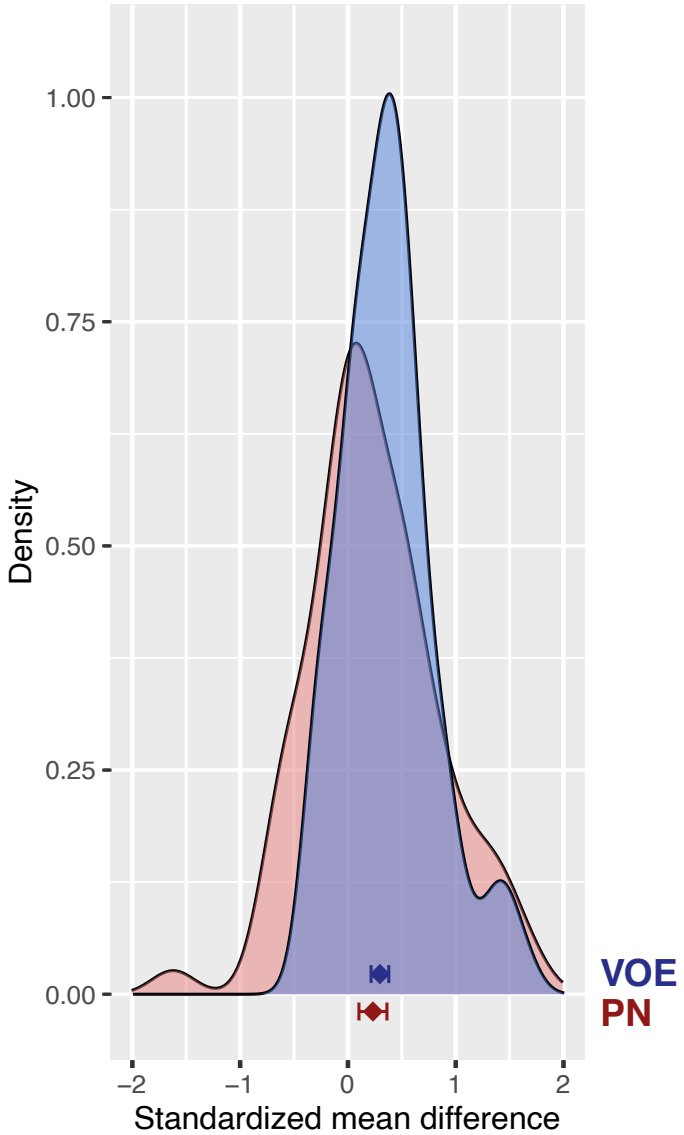
https://doi.org/10.1038/s41562-024-01965-x



From: Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71. doi: 10.1136/bmj.n71. For more information, visit: http://www.prisma-statement.org/

Extended Data Fig. 7 | Full PRISMA diagram for the current research. Thirty-three papers (76 studies) were included in the final analysis. We excluded one outlier paper (2 studies) that passed our screening process due to its extremely low variance relative to the other studies, which skewed some of the supplemental meta-analytic results. Our primary conclusions hold regardless of whether this study is included; see SI for details. Template retrieved from http://www.prisma-statement.org/PRISMAStatement/FlowDiagram.

https://doi.org/10.1038/s41562-024-01965-x



Extended Data Fig. 8 | Distribution of PN and VOE effects. Density plot over effect sizes in standard mean differences (N = 76 studies). Meta-analytic estimate of each effect with its 95% confidence interval are shown at the bottom of the plot.

nature portfolio

Corresponding author(s): Shari Liu

Last updated by author(s): Jul 1, 2024

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	For estimating values from published work from figures, we used WebPlotDigitizer version 4.5 (https://automeris.io/WebPlotDigitizer/). We also used Google Scholar, Proquest, and PubMed to search for papers. No other code was used to collect the data, but we did use resources and tutorials from MetaLab (https://langcog.github.io/metalab/).
Data analysis	We used R version 4.3.2 (2023-10-31) and the following R libraries: -tidyverse_2.0.0: data wrangling -ggplot2_3.4.4: visualization -metafor_4.4-0: meta-analysis -weightr_2.0.2: publication bias selection models -sjPlot_2.8.15: plot models -lme4_1.1-34: mega-analysis -lmerTest_3.1-3: mega-analysis -performance_0.10.8: check model assumptions -MASS_7.3-60: how data fits log and normal distribution -patchwork_1.1.3: create figures -viridis_0.6.4: create figures -pwr_1.3-0: power analysis reported in SI

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

All anonymized data associated with this paper are openly available at https://osf.io/b59km/.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	*** The current research does not involve interacting with or collecting new data from human subjects, and thus does not meet the definition of human subjects research. We nevertheless provide the information below, because the research involves using anonymized already-collected data from human subjects.*** We did not consider sex and gender as a covariate in our pre-registered analyses. However, we do report the % of infants whose sex assigned at birth was female, by parental report, for datasets that provided this information.
Reporting on race, ethnicity, or other socially relevant groupings	In the datasets we acquired, authors rarely reported information like race, ethnicity, or other demographic information other than sex; therefore, we could not do this either. We do report information about age and sex, which authors reported in their original research. We include infant age (averaged by condition, and individual infants' ages) as a covariate in many of our analyses; these values were calculated based on parent-reported birth dates)
Population characteristics	Our search criteria specified that participants should be typically developing infants, aged 3-12 months. See above.
Recruitment	No new data were collected, so there was no recruitment process involved in this research.
Ethics oversight	Each author group that contributed data collected these data with the approval of their local university IRB board.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	Our goal was to estimate the violation of expectation (VOE) effect and the perceptual novelty (PN) effect in the same experiments and datasets. We also wanted to focus the current research on two of the oldest topics in infant cognition (understanding of agents and objects), in a sample of studies containing enough variability (e.g. task domain; age of infants) for us to estimate moderator effects. Thus we conducted a systematic literature review, specifying the following inclusion criteria: (1) All literature, journal papers, theses, proceedings papers, and unpublished datasets after 1985, (2) that tested typically developing infants between 3 and 12 months of age, (3) on expectations about solid objects, or single agents engaging in intentional action, and (4) that employed an experimental design similar to those in Figure 1. Specifically, in order to be included: there had to be at least two habituation or familiarization trials before test trials; the expected and unexpected events had to be equally perceptually novel, relative to the previous trials, or the expected event had to be more perceptually novel than the unexpected event; and if the maximum duration of the familiarization and test trials differed, then the former had to be fairly long (at least 30s), or the ratio between maximum durations had to exceed 0.8.
Research sample	We used datasets from prior research on typically developing human infants, 3-12 months of age. Many authors did not report demographic information, and these studies likely follow the past trends of the field, focusing on predominantly White populations from North America and Western Europe.
Sampling strategy	This is a meta-analysis, for which the sample sizes are substantially larger than those from individual studies, and which offers us a rare opportunity to have high-powered samples (N = 1899, vs 20-30). We did not pre-register or specify the sample size of studies

	ahead of the research; instead we pre-registered the inclusion and exclusion criteria for the studies, knowing that we'd have a large sample size by virtue of the method.
Data collection	Data were extracted from published papers, or were sent to us by the authors. We found that 74.3% of studies reported that the data was generated by a naive human coder; 23.9% of studies did not specify whether the human coder was naive, and the remaining 1.8% were eye-tracking studies that did not use human coding. We note that this percentage is a conservative estimate, because we coded "yes" for this feature only if the authors explicitly mentioned the naiveness of the human rater. Some papers reported methods that make it likely that coders were naive (e.g. they saw a camera feed of the infant's face from another room; they looked through a peephole through the puppet stage at the infant), but did not explicitly mention experimenter or observer masking or blinding. We found that 93.4% of studies had a second coder check the reliability of the data, and reported information about interrater reliability; 4.6% of studies did not report this information, and the remaining 1.8% were eye-tracking studies that did not use human coding. Information about each study is shown in Extended Data Table 1
Timing	Our search began July 2022 and concluded November 2022.
Data exclusions	We screened a total of 2798 records, and in the end included 33 papers in our analyses, excluding 2597 records based on the title and abstract, and 167 based on article contents or missing key data. This resulted in 76 studies that were the focus of our analyses (experimental conditions), including data from individual infants in 60 of these studies. Additionally, we excluded one outlier paper (2 studies) that passed our screening process due to its extremely low variance relative to the other studies, which skewed some of the supplemental meta-analytic results. Our primary conclusions hold regardless of whether this study is included; see SI for details
Non-participation	No new data were collected for this research; we did not analyze the number of excluded participants from past datasets for the current research.
Randomization	No new data were collected for this research; for studies from prior research, researchers counterbalanced the order of test trials (expected first or unexpected first) across infants, either through pure random assignment (e.g. flipping a coin, running a pre-allocated number of infants per condition in a random order), or through stratefied random assignment (e.g. similar to pure random assignment, except that researchers tried to ensure that sex and age were balanced across counterbalancing orders).

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks	*Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.*
Novel plant genotypes	*Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.*
Authentication	*Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.*