

Using the past to estimate sensory uncertainty

Ulrik Beierholm^{1+*}, Tim Rohe^{2,3,4+}, Ambra Ferrari⁵, Oliver Stegle^{6,7,8}, Uta Noppeney^{2,5,9}

¹Psychology Department, Durham University, Durham, United Kingdom.

²Max Planck Institute for Biological Cybernetics, Tübingen, Germany

³Department of Psychiatry and Psychotherapy, University of Tübingen, Tübingen, Germany

⁴Department of Psychology, Friedrich-Alexander University Erlangen-Nuernberg, Erlangen, Germany

⁵Centre for Computational Neuroscience and Cognitive Robotics, University of Birmingham, Birmingham, United Kingdom.

⁶Max Planck Institute for Intelligent Systems, Tübingen, Germany

⁷European Molecular Biology Laboratory, Genome Biology Unit, Heidelberg, Germany

⁸Division for Computational Genomics & Systems Genetics, German Cancer Research Center, Heidelberg, Germany

⁹ Donders Institute for Brain, Cognition and Behaviour, Radboud University, Nijmegen, Netherlands

⁺ These authors contributed equally to this work.

^{*}Correspondence should be addressed to ulrik.beierholm@durham.ac.uk

26

Abstract

27 To form a more reliable percept of the environment, the brain needs to estimate its own sensory
28 uncertainty. Current theories of perceptual inference assume that the brain computes sensory
29 uncertainty instantaneously and independently for each stimulus. We evaluated this assumption
30 in four psychophysical experiments, in which human observers localized auditory signals that
31 were presented synchronously with spatially disparate visual signals. Critically, the visual noise
32 changed dynamically over time continuously or with intermittent jumps. Our results show that
33 observers integrate audiovisual inputs weighted by sensory uncertainty estimates that combine
34 information from past and current signals consistent with an optimal Bayesian learner that can
35 be approximated by exponential discounting. Our results challenge leading models of
36 perceptual inference where sensory uncertainty estimates depend only on the current stimulus.
37 They demonstrate that the brain capitalizes on the temporal dynamics of the external world and
38 estimates sensory uncertainty by combining past experiences with new incoming sensory
39 signals.

40

41

Introduction

42
43 Perception has been described as a process of statistical inference based on noisy sensory inputs
44 (Knill and Pouget, 2004, Knill and Richards, 1996). Key to this perceptual inference is the
45 estimation and/or representation of sensory uncertainty. Most prominently, in multisensory
46 perception a more reliable or ‘Bayes-optimal’ percept is obtained by integrating sensory signals
47 that come from a common source weighted by their relative reliabilities (i.e., precision or
48 inverse of variance) with less weight assigned to less reliable signals. Likewise, sensory
49 uncertainty shapes observers’ causal inference. It influences whether observers infer that signals
50 come from a common cause and should hence be integrated or else be processed independently
51 (Aller and Noppeney, 2019, Kording et al., 2007, Rohe et al., 2019, Rohe and Noppeney, 2015b,
52 Rohe and Noppeney, 2015a, Rohe and Noppeney, 2016, Wozny et al., 2010, Acerbi et al.,
53 2018). Indeed, accumulating evidence suggests that human observers are close to optimal in
54 many perceptual tasks (though see (Acerbi et al., 2014, Drugowitsch et al., 2016, Shen and Ma,
55 2016, Meijer et al., 2019)) and weight signals approximately according to their sensory
56 reliabilities (Alais and Burr, 2004, Ernst and Banks, 2002, Jacobs, 1999, Knill and Pouget,
57 2004, van Beers et al., 1999, Drugowitsch et al., 2014, Hou et al., 2019).

58 An unresolved question is how human observers compute their sensory uncertainty.
59 Current theories and experimental approaches generally assume that observers access sensory
60 uncertainty near-instantaneously and independently across briefly (≤ 200 ms) presented stimuli
61 (Ma and Jazayeri, 2014, Zemel et al., 1998). At the neural level, theories of probabilistic
62 population coding have suggested that sensory uncertainty may be represented instantaneously
63 in the gain of the neuronal population response (Ma et al., 2006, Hou et al., 2019). Yet, in our
64 natural environment, sensory noise often evolves at slow timescales. For instance, visual noise
65 slowly varies when walking through a snow storm. Observers may capitalize on the temporal
66 dynamics of the external world and use the past to inform current estimates of sensory
67 uncertainty. In this alternative account, more reliable estimates of sensory uncertainty would be
68 obtained by combining past estimates with current sensory inputs as predicted by Bayesian
69 learning.

70 To arbitrate between these two critical hypotheses, we presented observers with
71 audiovisual signals in synchrony but with a small spatial disparity in a sound localization task.
72 Critically, the spatial standard deviation (STD) of the visual signal changed dynamically over
73 time continuously (experiment 1-3) or discontinuously (i.e. with intermittent jumps; experiment
74 4). First, we investigated whether the influence of the visual signal location on observers’
75 perceived sound location depended on the noise only of the current visual signal or also of past

76 visual signals. Second, using computational modeling and Bayesian model comparison, we
77 formally assessed whether observers update their visual uncertainty estimates consistent with i.
78 an instantaneous learner, ii. an optimal Bayesian learner or iii. an exponential learner.
79

Results

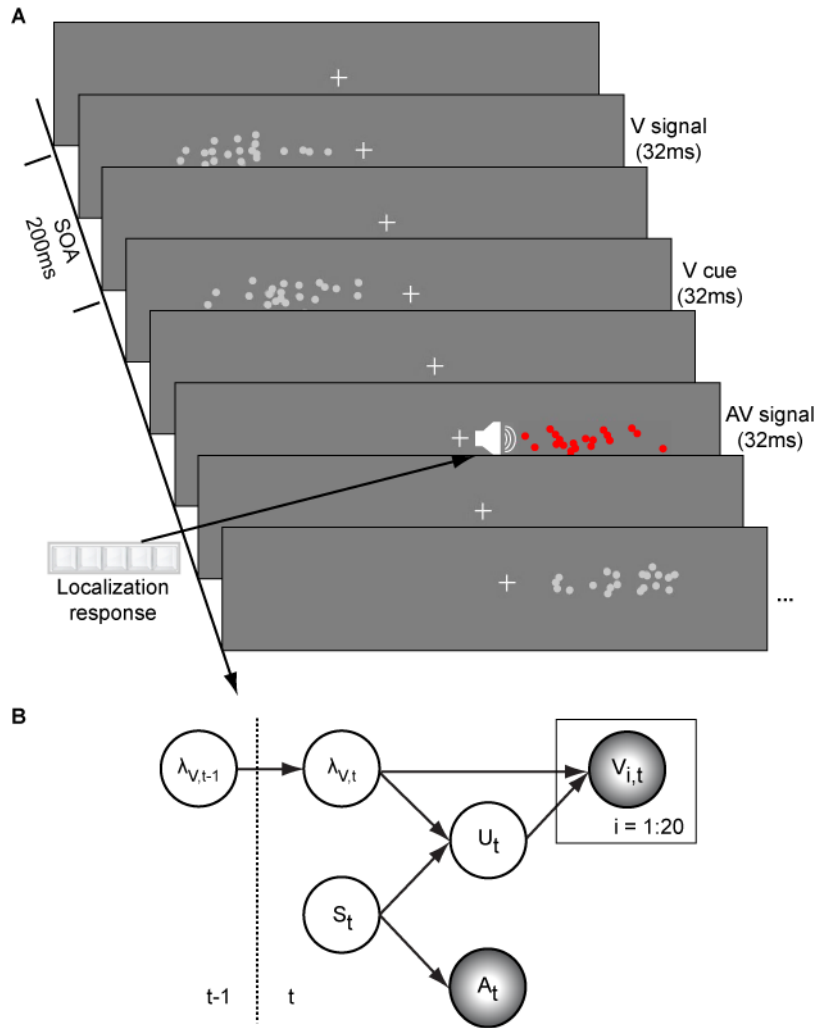


Figure 1. Audiovisual localization paradigm and Bayesian causal inference model for learning visual reliability. (A) Visual (V) signals (cloud of 20 bright dots) were presented every 200 ms for 32 ms. The cloud's location mean was temporally independently resampled from five possible locations (-10° , -5° , 0° , 5° , 10°) with an inter-trial asynchrony jittered between 1.4 and 2.8 s. In synchrony with the change in the cloud's mean location, the dots changed their colour and a sound was presented (AV signal) which the participants localized using five response buttons. The location of the sound was sampled from the two possible locations adjacent to the visual cloud's mean location (i.e. $\pm 5^\circ$ AV spatial discrepancy). (B) The generative model for the Bayesian learner explicitly modelled the potential causal structures, i.e. whether visual (V_i) signals and an auditory (A) signal were generated by one common audiovisual source S_t , i.e. $C = 1$, or by two independent sources S_{Vt} and S_{At} , i.e. $C = 2$ (n.b. only the model component for the common source case is shown to illustrate the temporal updating, for complete generative model, see Figure 1-figure supplement 1). Importantly, the reliability (i.e., $1/\text{variance}$) of the visual signal at time t (λ_t) depends on the reliability of the previous visual signal (λ_{t-1}) for both model components (i.e. common and independent sources).

81 In a spatial localization task, we presented participants with audiovisual signals in a series of
 82 four experiments, in which the physical visual noise changed dynamically over time either
 83 continuously or discontinuously (Figure 1). Visual (V) signals (clouds of 20 bright dots) were
 84 presented every 200 ms for a duration of 32 ms. The cloud's horizontal standard deviation
 85 (STD) varied over time at this temporal rate of 5 Hz either continuously (experiments 1-3) or

86 discontinuously with intermittent jumps (experiment 4). The cloud's location mean was
 87 temporally independently resampled from five possible locations (-10° , -5° , 0° , 5° , 10°) on each
 88 trial with the inter-trial asynchrony jittered between 1.4 and 2.8 s. In synchrony with the change
 89 in the cloud's mean location, the dots changed their colour and a sound was presented (AV
 90 signal). The location of the sound was sampled from the two possible locations adjacent to the
 91 visual cloud's mean location (i.e. $\pm 5^\circ$ AV spatial disparity). Participants localized the sound
 92 and indicated their response using five response buttons.

93 The small audiovisual disparity enabled an influence of the visual signal location on the
 94 perceived sound location as a function of visual noise (Alais and Burr, 2004, Battaglia et al.,
 95 2003, Meijer et al., 2019). As a result, observers' visual uncertainty estimate could be quantified
 96 in terms of the relative weight of the auditory signal on the perceived sound location with a
 97 greater auditory weight indicating that observers estimated a greater visual uncertainty.

98 In the first three experiments, we used continuous sequences, where the visual cloud's STD
 99 changed periodically according to a sinusoid ($n = 25$; period = 30 s), a random walk (RW1; $n = 33$;
 100 period = 120 s) or a smoothed random walk (RW2; $n = 19$; period = 30 s; Figure 2). In an
 101 additional fourth experiment, we inserted abrupt increases or decreases into a sinusoidal
 102 evolution of the visual cloud's STD ($n = 18$, period = 30 s, Figure 5). We will first describe the
 103 results for the three continuous sequences followed by the discontinuous sequence.

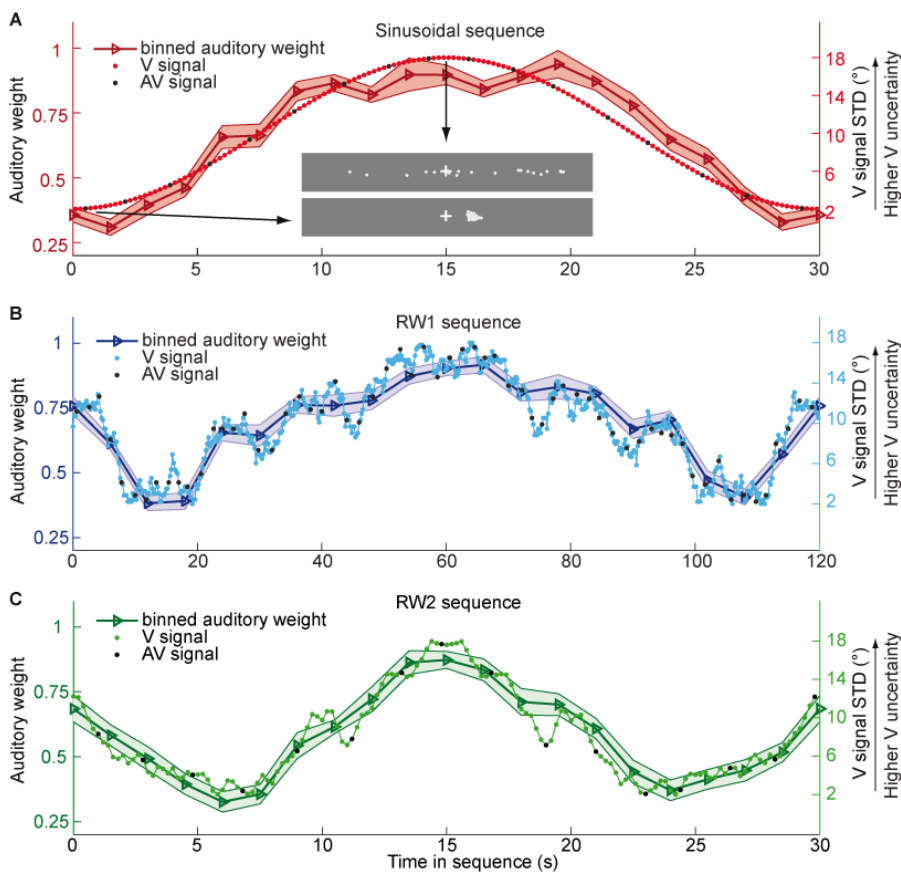


Figure 2. Time course of visual noise and relative auditory weights for continuous sequences of visual noise. The visual noise (i.e., STD of the cloud of dots, right ordinate) and the relative auditory weights (mean across participants \pm SEM, left ordinate) are displayed as a function of time. The STD of the visual cloud was manipulated as (A) a sinusoidal (period 30s, N = 25), (B) a random walk (RW1, period 120s, N = 33) and (C) a smoothed random walk (RW2, period 30s, N = 19). The period for the RW1 sequence is 120 s, while the periods of the sinusoidal and RW2 is only 30 s. The overall dynamics as quantified by the power spectrum is faster for RW2 than RW1 (peak in frequency range [0 0.2] Hz: Sinusoid: 0.033 Hz, RW1: 0.025 Hz, RW2: 0.066 Hz). The RW1 and RW2 sequences were mirror-symmetric around the half-time (i.e., the second half was the reversed first half). The visual clouds were re-displayed every 200 ms (i.e., at 5 Hz). The trial onsets, i.e. audiovisual (AV) signals (color change with sound presentation, black dots), were interspersed with an inter-trial asynchrony jittered between 1.4 and 2.8 s. On each trial observers located the sound. The relative auditory weights were computed based on regression models for the sound localization responses separately for each of the 20 temporally adjacent bins that cover the entire period within each participant. The relative auditory weights vary between one (i.e. pure auditory influence on the localization responses) and zero (i.e. pure visual influence). For illustration purposes, the cloud of dots for the lowest (i.e., V signal STD = 2°) and the highest (i.e., V signal STD = 18°) visual variance are shown in (A).

104 We assigned the sound localization responses and the associated physical visual noise
 105 (i.e., the cloud's STD) to 20 (resp. 15 for experiment 4) temporally adjacent bins covering the
 106 entire period of each of the three sequences. Each experiment repeated the same 30 s (Sin,
 107 RW2) or 60s (RW1) period throughout the experiment resulting in \sim 32 periods for the RW2
 108 and \sim 130 periods for the Sin and RW1 sequences. The trial and hence sound onsets were jittered
 109 with respect to this periodic evolution of the visual cloud's STD resulting in a greater effective
 110 sampling rate than expected for an inter-trial asynchrony of 1.4 – 2.8 s. In total, we assigned at
 111 least 44-87 trials to each bin (Supplementary file 1-Table 1). We quantified the auditory and
 112 visual influence on observers' perceived auditory location for each bin based on regression
 113 models (separately for each of the 20 temporally adjacent bins). For instance, for bin = 1 we
 114 computed:

$$115 \quad (1) R_{A,trial,bin=1} = L_{A,trial,bin=1} \beta_{A,bin=1} + L_{V,trial,bin=1} \beta_{V,bin=1} + \beta_{const,bin=1} +$$

$$116 \quad e_{trial,bin=1}$$

117 with $R_{A,trial,bin=1}$ = Localization response for trial t and bin 1; $L_{A,trial,bin=1}$ or $L_{V,trial,bin=1}$ =
 118 'true' auditory or visual location for trial t and bin 1; $\beta_{A,bin=1}$ or $\beta_{V,bin=1}$ = auditory or visual
 119 weight for bin 1; $\beta_{const,bin=1}$ = constant term; $e_{trial,bin=1}$ = error term. For each bin b, we thus
 120 obtained one auditory and one visual weight estimate. The relative auditory weight for a
 121 particular bin was computed as $w_{A,bin} = \beta_{A,bin} / (\beta_{A,bin} + \beta_{V,bin})$.

122 Figure 2 and Figure 3 show the temporal evolution of the STD of the physical visual
 123 noise and observers' relative auditory weight indices $w_{A,bin}$. If observers estimate sensory
 124 uncertainty instantaneously, observer's relative auditory weight indices should closely track the
 125 visual cloud's STD (Figure 2). By contrast, we observed systematic biases: while the temporal

126 evolution of the physical visual noise was designed to be symmetrical for each time period, we
 127 observed a temporal asymmetry for w_A in all of the three experiments. For the monotonic
 128 sinusoidal sequence, w_A was smaller for the 1st half of each period, when visual noise increased,
 129 than the 2nd half, when visual noise decreased over time (Figure 3A). For the non-monotonic
 130 RW1 and RW2 sequences, we observed more complex temporal profiles, because the visual
 131 noise increased and decreased in each half. w_A was larger for increasing visual noise in the 1st
 132 as compared to the 2nd half, while w_A was smaller for decreasing visual noise in the 1st as
 133 compared to the 2nd half (Figure 3B, C). These impressions were confirmed statistically in 2
 134 (1st vs. flipped 2nd half) \times 9 (bins) repeated measures ANOVAs (Table 1) showing a significant
 135 main effect of the 1st versus flipped 2nd half period for the sinusoidal ($F(1, 24) = 12.162$, $p =$
 136 0.002 , partial $\eta^2 = 0.336$) and the RW1 sequence ($F(1, 32) = 14.129$, $p < 0.001$, partial $\eta^2 =$
 137 0.306). For the RW2 sequence, we observed a significant interaction ($F(4.6, 82.9) = 3.385$, $p =$
 138 0.010 , partial $\eta^2 = 0.158$), because the visual noise did not change monotonically within each
 139 half period. Instead, monotonic increases and decreases in visual noise alternated at nearly the
 140 double frequency in RW2 as compared to RW1. The asymmetry in the auditory weights' time
 141 course across the three experiments suggested that the visual noise in the past influenced
 142 observers' current visual uncertainty estimate resulting in smaller auditory weights for
 143 ascending visual noise and greater auditory weights for descending visual noise.

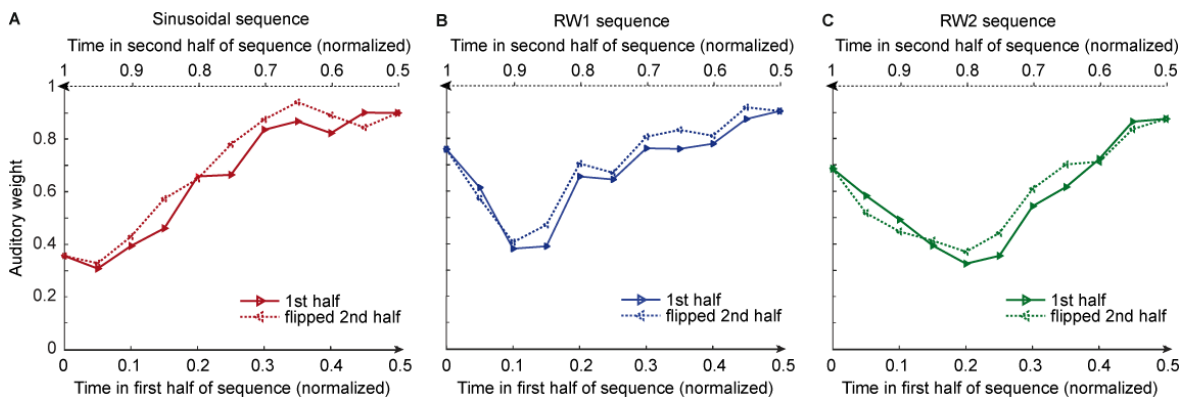


Figure 3. Observers' relative auditory weights for continuous sequences of visual noise. Relative auditory weights w_A of the 1st (solid) and the flipped 2nd half (dashed) of a period (binned into 20 bins) plotted as a function of the normalized time in the sinusoidal (red), the RW1 (blue) and the RW2 (green) sequences. Relative auditory weights were computed from auditory localization responses of human observers.

144 To further investigate the influence of past visual noise on observers' auditory weights,
 145 we estimated a regression model in which the relative auditory weights w_A for each of the 20
 146 bins were predicted by the visual STD in the current bin and the difference in STD between the
 147 current and the previous bin (see equation (2)). Indeed, both the current visual STD ($p < 0.001$
 148 for all three sequences; Sinusoid: $t(24)=15.767$, Cohen's $d=3.153$; RW1: $t(32)= 15.907$,

Table 1. Analyses of the temporal asymmetry of the relative auditory weights across the four sequences of visual noise using repeated measures ANOVAs with the factors sequence part (1st vs. flipped 2nd half), bin and jump position (only for the sinusoidal sequences with intermittent jumps).

	Effect	F	df1	df2	p	Partial η^2
Sinusoid	Part	12.162	1	24	0.002	0.336
	Bin	92.007	3.108	74.584	<0.001	0.793
	PartXBin	2.167	2.942	70.617	0.101	0.083
RW1	Part	14.129	1	32	0.001	0.306
	Bin	76.055	4.911	157.151	<0.001	0.704
	PartXBin	1.225	4.874	155.971	0.300	0.037
RW2	Part	2.884	1	18	0.107	0.138
	Bin	60.142	3.304	59.467	<0.001	0.770
	PartXBin	3.385	4.603	82.849	0.010	0.158
Sinusoid with intermittent jumps	Jump	28.306	2	34	<0.001	0.625
	Part	24.824	1	17	<0.001	0.594
	Bin	76.476	1.873	31.839	<0.001	0.818
	JumpXPart	0.300	2	34	0.743	0.017
	JumpXBin	8.383	3.309	56.247	<0.001	0.330
	PartXBin	1.641	3.248	55.222	0.187	0.088
	JumpXPartXBin	0.640	5.716	97.175	0.690	0.036

Note: The factor bin comprised 9 levels in the first three and 7 levels in the fourth sequence. In this sequence, the factor Jump comprised three levels. If Mauchly tests indicated significant deviations from sphericity ($p < 0.05$), we report Greenhouse-Geisser corrected degrees of freedom and p values.

149 Cohen's $d=2.769$; RW2: $t(18)=12.978$, Cohen's $d=2.977$, two sided one-sample t test against
 150 zero) and the difference in STD between the current and the previous bin (i.e. Sinusoid $t(24)=$
 151 -3.687 , $p = 0.001$, Cohen's $d=-0.737$; RW1 $t(32)= -2.593$, $p = 0.014$, Cohen's $d=-0.451$; RW2
 152 $t(18)=-2.395$, $p = 0.028$, Cohen's $d=-0.549$) significantly predicted observers' relative auditory
 153 weights (for complementary results of nested model comparison see Appendix 1 and
 154 Supplementary file 1-Table 5). Collectively, these results suggest that observers' visual

155 uncertainty estimates (as indexed by the relative auditory weights w_A) depend not only on the
 156 current sensory signal, but also on the recent history of the sensory noise. These results were
 157 also validated in a control analysis that regressed out and thus accounted for potential influences
 158 of the previous visual location on observers' sound localization, suggesting that the effects of
 159 past visual uncertainty cannot be explained by effects of past visual location mean (Appendix
 160 1, Figure 2-figure supplement 1, Supplementary file 1-Table 2-4).

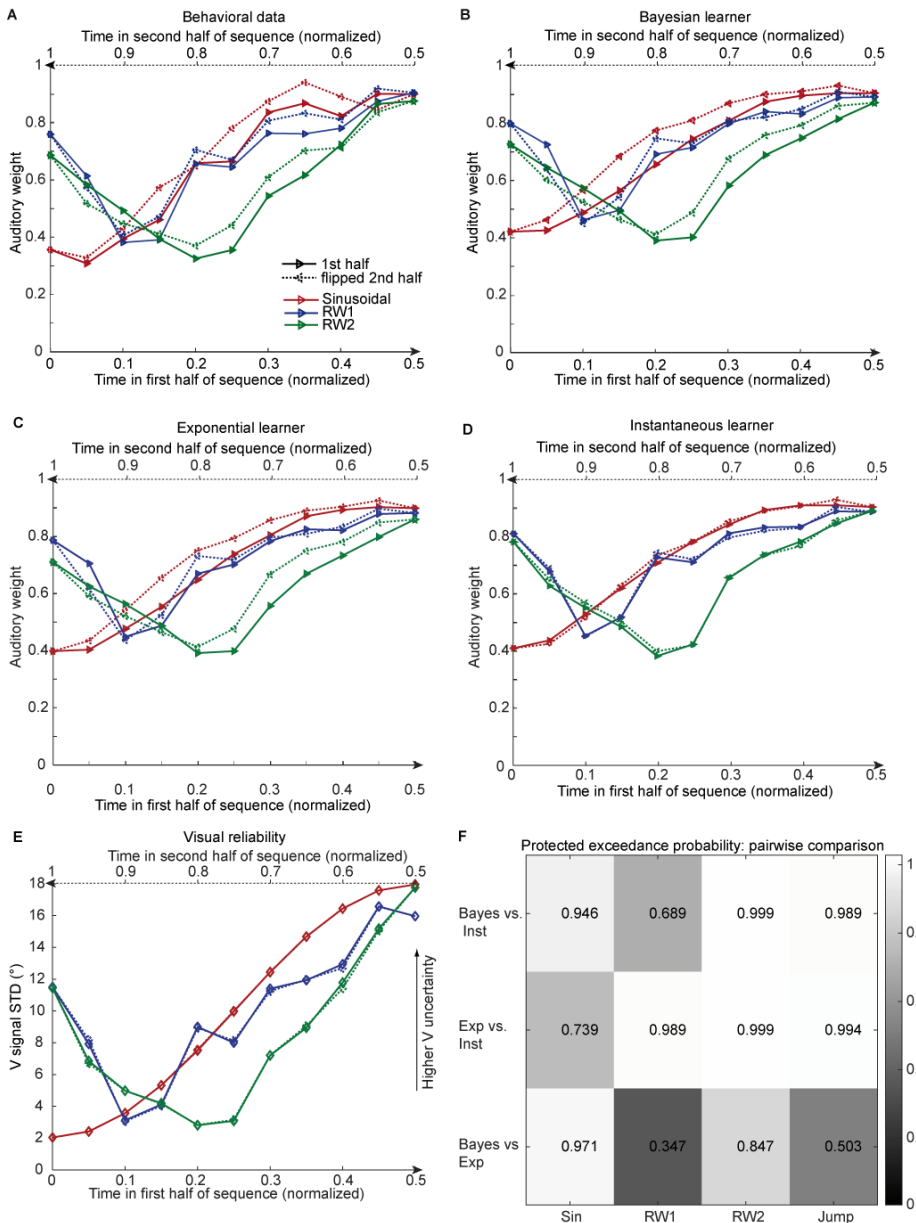


Figure 4. Observed and predicted relative auditory weights for continuous sequences of visual noise. Relative auditory weights w_A of the 1st (solid) and the flipped 2nd half (dashed) of a period (binned into 20 bins) plotted as a function of the normalized time in the sinusoidal (red), the RW1 (blue) and the RW2 (green) sequences. Relative auditory weights were computed from auditory localization responses of human observers (A), Bayesian (B), exponential (C) or instantaneous (D) learning models. For comparison, the standard deviation of the visual signal is shown in (E). Please note that all models were fitted to observers' auditory localization responses (i.e. not the auditory weight w_A). (F) Bayesian model comparison – Random effects analysis: The matrix shows the protected exceedance probability (color coded and indicated by the numbers) for pairwise

comparisons of the Instantaneous (Inst), Bayesian (Bayes) and Exponential (Exp) learners separately for each of the four experiments. Across all experiments we observed that the Bayesian or the Exponential learner outperformed the Instantaneous learner (i.e. a protected exceedance probability > 0.94) indicating that observers used the past to estimate sensory uncertainty. However, it was not possible to arbitrate reliably between the Exponential and the Bayesian learner across all experiments (c.f. protected exceedance probability in bottom row).

161 To characterize how human observers use information from the past to estimate current
 162 sensory uncertainty, we compared three computational models that differed in how visual
 163 uncertainty is learnt over time (Figure 4): Model 1, the instantaneous learner, estimates visual
 164 uncertainty independently for each trial as assumed by current standard models. Model 2, the
 165 optimal Bayesian learner, estimates visual uncertainty by updating the prior uncertainty
 166 estimate obtained from past visual signals with the uncertainty estimate from the current signal.
 167 Model 3, the exponential learner, estimates visual uncertainty by exponentially discounting past
 168 uncertainty estimates. All three models account for observers' uncertainty about whether
 169 auditory and visual signals were generated by common or independent sources by explicitly
 170 modeling the two potential causal structures (Kording et al., 2007) underlying the audiovisual
 171 signals (n.b. only the model component pertaining to the 'common cause' case is shown in
 172 Figure 1B, for the full model see Figure 1-figure supplement 1). Models were fit individually
 173 to observers' data by sampling from the posterior over parameters for each observer (Table 2).

Table 2. Model parameters (median), absolute WAIC and relative Δ WAIC values for the three candidate models in the four sequences of visual noise.

Sequence	Model	σ_A	P_{common}	σ_0	κ or γ	WAIC	Δ WAIC
Sinusoid	Instantaneous learner	5.56	0.63	8.95	-	81931.2	109.9
	Bayesian learner	5.64	0.65	9.03	κ : 7.37	81821.3	0
	Exponential discounting	5.62	0.64	9.02	γ : 0.23	81866.9	45.6
RW1	Instantaneous learner	6.30	0.69	8.46	-	110051.2	89.0
	Bayesian learner	6.29	0.72	8.68	κ : 8.06	109962.2	0
	Exponential discounting	6.26	0.70	8.75	γ : 0.33	109929.9	-32.3
RW2	Instantaneous learner	6.36	0.72	10.79	-	62576.4	201.3
	Bayesian learner	6.49	0.78	10.9	κ : 6.7	62375.2	0
	Exponential discounting	6.46	0.73	11.0	γ : 0.25	62421.5	46.3
Sinusoid with intermittent jumps	Instantaneous learner	6.38	0.65	8.19	-	83891.4	94.9
	Bayesian learner	6.45	0.68	8.26	κ : 6.13	83796.5	0
	Exponential discounting	6.43	0.67	8.20	γ : 0.24	83798.1	1.64

Note: WAIC values were computed for each participant and summed across participants. A low WAIC indicates a better model. Δ WAIC is relative to the WAIC of the Bayesian learner.

174 We compared the three models in a fixed and random effects analysis (Penny et al.,
175 2010, Rigoux et al., 2014) using the Watanabe-Akaike information criterion (WAIC) as
176 appropriate for evaluating model samples (Gelman et al., 2014) (i.e., a low WAIC indicates a
177 better model, a difference greater than 10 is considered very strong evidence for a model). In
178 the fixed-effects analysis (see Table 2 for details), the Bayesian learner was substantially better
179 than the instantaneous learner across all three experiments, but outperformed the exponential
180 learner reliably only in the sinusoidal sequence. Likewise, the random-effects analysis based
181 on hierarchical Bayesian model selection (Penny et al., 2010, Rigoux et al., 2014) showed a
182 protected exceedance probability that was substantially greater for the Bayesian learner (Sin,
183 RW2) or the exponential learner (RW1, RW2) than for the instantaneous learner (Figure 4F).
184 However, the direct comparison between the Bayesian and the exponential learner did not
185 provide consistent results across experiments. As shown in Figure 4 A and B, both the Bayesian
186 and the exponential learner accurately reproduced the temporal asymmetry for the auditory
187 weights across all three experiments.

188 From the optimal Bayesian learner we inferred observers' estimated rate of change in
189 visual reliability (i.e. parameter $1/\kappa$). The sinusoidal sequence was estimated to change at a
190 faster pace (median $\kappa = 7.4$ across observers, 95 percent confidence interval, 95%CI [4.8, 10.8]
191 estimated via bootstrapping) than the RW1 sequence (median $\kappa = 8.1$, 95%CI [7.0,14.9]), but
192 slower than the RW2 sequence (median $\kappa = 6.7$, 95%CI [4.4,11.2]) indicating that the Bayesian
193 learner accurately inferred that visual reliability changed at different pace across the three
194 continuous sequences (see legend of Figure 2). Likewise, the learning rates $1-\gamma$ of the
195 exponential learner accurately reflect the different rates of change across the sequences
196 (Sinusoid $\gamma = 0.23$, 95%CI [0.14, 0.28]; RW1: $\gamma = 0.33$, 95%CI [0.21, 0.38]; RW2: $\gamma = 0.25$,
197 95%CI [0.21, 0.29]). Both the Bayesian and the exponential learner thus estimated a smaller
198 rate of change for the RW1 than for the sinusoidal sequence – though caution needs to be
199 applied when interpreting these results given the extensive confidence intervals. Further, the
200 learning rates of the exponential learner imply that observers gave the visual inputs presented
201 4.1 (Sinusoid), 5.4 (RW1) and 4.3 (RW2) seconds before the current stimulus 5% of the weight
202 they assigned to the current visual input to estimate the visual reliability.

203 To further disambiguate between the Bayesian and the exponential learner, we designed
204 a fourth experimental 'jump sequence' that introduced abrupt increases or decreases in physical
205 visual noise at three positions into the sinusoidal sequence (Figure 5A). Using the same analysis

206 approach as for experiments 1-3, we replicated the temporal asymmetry for the auditory weights
 207 (Figure 5B). For all three ‘jump positions’ w_A was significantly smaller for the 1st half of each
 208 period, when visual noise increased, than the 2nd half, when visual noise decreased over time.
 209 The 3 (jump positions) x 2 (1st vs. flipped 2nd half) x 7 (bins) repeated measures ANOVA
 210 showed a significant main effect of 1st versus flipped 2nd period’s half ($F(1,17) = 24.824$, $p <$
 211 0.001 , partial $\eta^2 = 0.594$), while this factor was not involved in any higher-order interaction
 212 (see Table 1). Further, in a regression model the current visual STD ($t(17) = 11.655$, $p < 0.001$,
 213 Cohen’s $d = 2.747$) and the difference between current and previous STD ($t(17) = -4.768$, $p <$
 214 0.001 , Cohen’s $d = -1.124$) significantly predicted the relative auditory weights. Thus, we
 215 replicated our finding that the visual noise in the past influenced observers’ current visual
 216 uncertainty estimate as indexed by the relative auditory weights w_A .

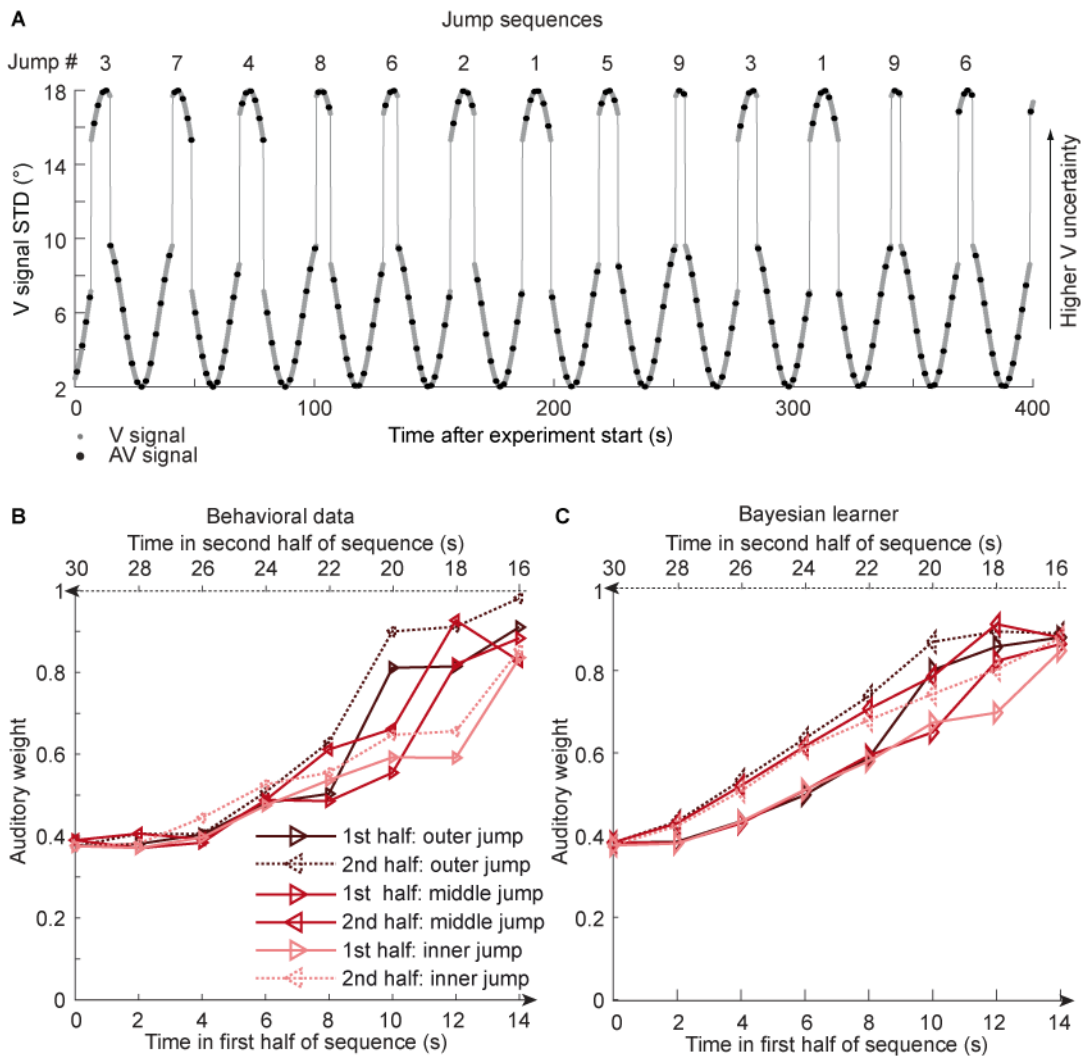


Figure 5. Time course of visual noise and relative auditory weights for sinusoidal sequence with intermittent jumps in visual noise (N = 18). (A) The visual noise (i.e., STD of the cloud of dots, right ordinate) is displayed as a function of time. Each cycle included one abrupt increase and decrease in visual noise. The sequence of visual clouds was presented every 200 ms (i.e., at 5 Hz) while audiovisual (AV) signals (black dots) were interspersed with an inter-trial asynchrony jittered between 1.4 and 2.8 s. (B, C) Relative auditory weights w_A of the 1st (solid) and the flipped 2nd half (dashed) of a period (binned into 15 bins) plotted as a function of the time in the sinusoidal sequence with intermitted inner (light gray), middle (gray) and outer (dark gray) jumps. Relative auditory weights were computed from auditory localization responses of human observers (B) and the Bayesian learning model (C). Please note that all models were fitted to observers' auditory localization responses (i.e. not the auditory weight w_A).

217 Bayesian model comparison using a fixed-effects analysis showed that both the
218 Bayesian learner and the exponential learner substantially outperformed the instantaneous
219 learner (see Table 2). However, consistent with our Bayesian model comparison results for the
220 continuous sequences, the Bayesian learner did not provide a better explanation for observers'
221 responses than the exponential learner (Δ WAIC = +2, see Table 2, Figure 5C and Figure 5-
222 figure supplement 1A). Likewise, a random-effects analysis based on hierarchical Bayesian
223 model selection showed that the Bayesian and the exponential learners outperformed the
224 instantaneous learner, but again we were not able to adjudicate between the Bayesian and
225 exponential learner (Figure 4F, see also methods and results in Appendix 1, Figure 5-figure
226 supplement 2 and Supplementary file 1-Table 6 for further analyses justifying the choice of
227 continuous learning models in the jump sequence).

228 In summary, across four experiments that used continuous and discontinuous sequences
229 of visual noise, we have shown that the Bayesian or exponential learners outperform the
230 instantaneous learner. However, across the four experiments we were not able to decide whether
231 observers adapted to changes in visual noise according to a Bayesian or an exponential learner.
232 The key feature that distinguishes between the Bayesian and the exponential learner is that only
233 the Bayesian learner adapts dynamically based on its uncertainty about its visual reliability
234 estimates. As a consequence, the Bayesian learner should adapt faster than the exponential
235 learner to increases in physical visual noise (i.e. spread of the visual cloud) but slower to
236 decreases in visual noise. From the Bayesian learner's perspective, the faster learning for
237 increases in visual noise emerges because it is unlikely that visual dots form a large spread
238 cloud under the assumption that the true visual spread of the cloud is small. Conversely, the
239 Bayesian learner will adapt more slowly to decreases in visual variance, because under the
240 assumption of a visual cloud with a large spread visual dots may form a small cloud by chance.
241 Indeed, previous research has shown that observers adapt their variance estimates faster for
242 changes from small to large than for changes from large to small variance (Berniker et al.,
243 2010). However, these results have been shown for learning about a hidden variable such as the

244 prior that defines the spatial distribution from which an object's location is sampled. In our
245 study, we manipulated the variance of the likelihood, i.e. the variance of the clouds of dots.

246 Asymmetric differences in adaptation rate between the exponential and the Bayesian
247 learner should thus be amplified if we increase observer's uncertainty about its visual reliability
248 estimate by reducing the number of dots of the visual cloud from 20 to 5 dots. Based on
249 simulations, we therefore explored whether we could experimentally discriminate between the
250 Bayesian and exponential learner using continuous sinusoidal or discontinuous 'jump'
251 sequences with visual clouds of only 5 dots. For the two sequences, we simulated the sound
252 localization responses of 12 observers based on the Bayesian learner model and fitted the
253 Bayesian and exponential learner models to the responses of each simulated Bayesian observer.
254 Figure 6 shows observers' auditory weights indexing their estimated visual reliability across
255 time that we obtained from the fitted responses of the Bayesian (blue) and the exponential
256 learner (green). The simulations reveal the characteristic differences in how the Bayesian and
257 the exponential learner adapt their visual uncertainty estimates to increases and decreases in
258 visual noise. As expected, the Bayesian learner adapts its visual uncertainty estimates faster
259 than the exponential learner to increases in visual noise, but slower to decreases in visual noise.
260 Nevertheless, these differences are relatively small, so that the difference in mean log likelihood
261 between the Bayesian and exponential learner is only -1.82 for the sinusoidal sequence and -
262 2.74 for the jump sequence.

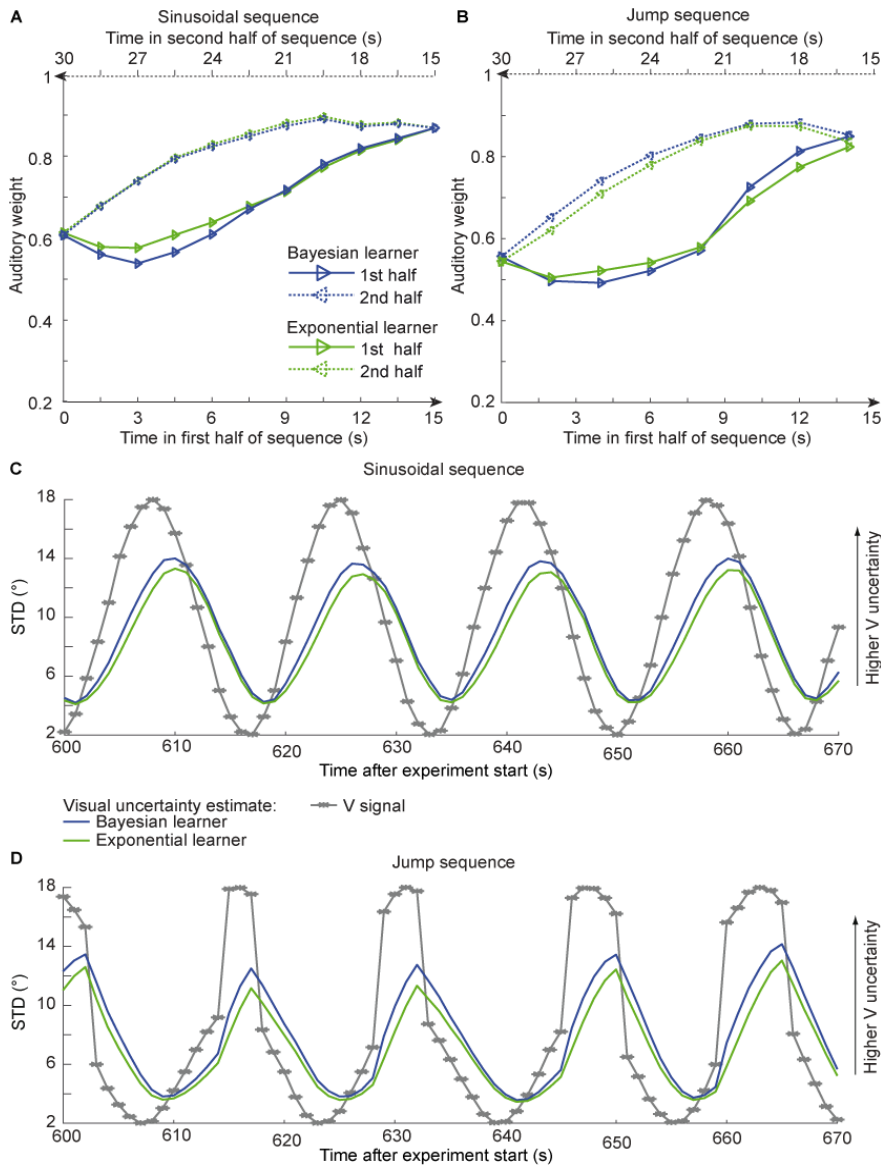


Figure 6. Time course of the relative auditory weights, the standard deviation of the visual cloud and the standard deviation of the visual uncertainty estimates. (A) Relative auditory weights w_A of the 1st (solid) and the flipped 2nd half (dashed) of a period (binned into 15 bins) plotted as a function of the time in the sinusoidal sequence. Relative auditory weights were computed from the predicted auditory localization responses of the Bayesian (blue) or exponential (green) learning models fitted to the simulated localization responses of a Bayesian learner based on visual clouds of 5 dots. (B) Relative auditory weights w_A computed as in (A) for the sinusoidal sequence with intermitted jumps. Only the outer-most jump (black in Figure 5B/C and Figure 5-figure supplement 1) is shown. (C, D) Standard deviation (STD) of the visual cloud of 5 dots (grey) and the STD of observers' visual uncertainty as estimated by the Bayesian (blue) and exponential (green) learners (that were fitted to the simulated localization responses of a Bayesian learner) as a function of time for the sinusoidal sequence (C) and in the sinusoidal sequence with intermitted jumps (D). Note that only an exemplary time course from 600-670 s after the experiment start is shown.

263 Next, we investigated whether our experiments successfully mimicked situations in
 264 which observers benefit from integrating past and current information to estimate their sensory
 265 uncertainty. We compared the accuracy of the instantaneous, exponential and Bayesian
 266 learner's visual uncertainty estimates in terms of their mean absolute deviation (in percentage)

267 from the true variance. For Gaussian clouds of 20 dots, the instantaneous learner's error in the
268 visual uncertainty estimates of 21.7 % is reduced to 13.7 % and 14.9% for the exponential and
269 Bayesian learners, respectively (with best fitted $\gamma = 0.6$, in the sinusoidal sequence). For
270 Gaussian clouds composed of only 5 dots, the exponential and Bayesian learners even cut down
271 the error by half (i.e. 46.8 % instantaneous learner, 29.5 % exponential learner, 23.9 % Bayesian
272 learner, with best fitted $\gamma = 0.7$).

273 Collectively, these simulation results suggest that even in situations in which observers
274 benefit from combining past with current sensory inputs to obtain more precise uncertainty
275 estimates, the exponential learner is a good approximation of the Bayesian learner, making it
276 challenging to dissociate the two experimentally based on noisy human behavioural responses.

277

278

Discussion

279
280 The results from our four experiments challenge classical models of perceptual inference where
281 a perceptual interpretation is obtained using a likelihood that depends solely on the current
282 sensory inputs (Ernst and Banks, 2002). These models implicitly assume that sensory
283 uncertainty (i.e., likelihood variance) is instantaneously and independently accessed from the
284 sensory signals on each trial based on initial calibration of the nervous system (Jacobs and Fine,
285 1999). Most prominently, in the field of cue combination it is generally assumed that sensory
286 signals are weighted by their uncertainties that are estimated only from the current sensory
287 signals (Alais and Burr, 2004, Ernst and Banks, 2002, Jacobs, 1999) (but see (Mikula et al.,
288 2018, Triesch et al., 2002)).

289 By contrast, our results demonstrate that human observers integrate inputs weighted by
290 uncertainties that are estimated jointly from past and current sensory signals. Across the three
291 continuous and the one discontinuous jump sequences, observers' current visual reliability
292 estimates were influenced by visual inputs that were presented 4-5 s in the past albeit their
293 influence amounted to only 5% of the current visual signals.

294 Critically, observers adapted their visual uncertainty estimates flexibly according to the rate of
295 change in the visual noise across the experiments. As predicted by both Bayesian and
296 exponential learning models, observers' visual reliability estimates relied more strongly on past
297 sensory inputs, when the visual noise changed more slowly across time. While observers did
298 not explicitly notice that each of the four experiments was composed of repetitions of
299 temporally symmetric sequence components, we cannot fully exclude that observers may have
300 implicitly learnt this underlying temporal structure. However, implicit or explicit knowledge of
301 this repetitive sequence structure should have given observers the ability to predict and preempt
302 future changes in visual reliability and therefore attenuated the temporal lag of the visual
303 reliability estimates. Put differently, our experimental choice of repeating the same sequence
304 component over and over again in the experiment cannot explain the influence of past signals
305 on observers' current reliability estimate, but should have reduced or even abolished it.

306 Importantly, the key feature that distinguishes the Bayesian from the exponential learner is how
307 the two learners adapt to increases versus decreases in visual noise. Only the Bayesian learner
308 represents and accounts for its uncertainty about its visual reliability estimates. As compared to
309 the exponential learner, it should therefore adapt faster to increases but slower to decreases in
310 visual noise (e.g. see Berniker et al. (2010)). Our simulation results show this profile
311 qualitatively, when the learner's uncertainty about its visual reliability estimate is increased by
312 reducing the number of dots (see Figure 6). But even for visual clouds of five dots, the

313 differences in learning curves between the Bayesian and exponential learner are very small
314 making it difficult to adjudicate between them given noisy observations from real observers.
315 Unsurprisingly, therefore, Bayesian model comparison showed consistently across all four
316 experiments that observers' localization responses can be explained equally well by an optimal
317 Bayesian and an exponential learner. These results converge with a recent study showing that
318 learning about a hidden variable such as observers' priors can be accounted for by an
319 exponential averaging model (Norton et al., 2019).

320 Collectively, our experimental and simulation results suggest that under circumstances where
321 observers substantially benefit from combining past and current sensory inputs for estimating
322 sensory uncertainty, optimal Bayesian learning can be approximated well by more simple
323 heuristic strategies of exponential discounting that update sensory weights with a fixed learning
324 rate irrespective of observers' uncertainty about their visual reliability estimate (Ma and
325 Jazayeri, 2014, Shen and Ma, 2016). Future research will need to assess whether observers
326 adapt their visual uncertainty estimates similarly if visual noise is manipulated via other
327 methods such as stimulus luminance, duration or blur.

328 From the perspective of neural coding, our findings suggest that current theories of probabilistic
329 population coding (Beck et al., 2008, Ma et al., 2006, Hou et al., 2019) may need to be extended
330 to accommodate additional influences of past experiences on neural representations of sensory
331 uncertainties. Alternatively, the brain may compute sensory uncertainty using strategies of
332 temporal sampling (Fiser et al., 2010).

333 In conclusion, our study demonstrates that human observers do not access sensory uncertainty
334 instantaneously from the current sensory signals alone, but learn sensory uncertainty over time
335 by combining past experiences and current sensory inputs as predicted by an optimal Bayesian
336 learner or approximate strategies of exponential discounting. This influence of past signals on
337 current sensory uncertainty estimates is likely to affect learning not only at slower timescales
338 across trials (i.e. as shown in this study), but also at faster timescales of evidence accumulation
339 within a trial (Drugowitsch et al., 2014). While our research unravels the impact of prior sensory
340 inputs on uncertainty estimation in a cue combination context, we expect that they reveal
341 fundamental principles of how the human brain computes and encodes sensory uncertainty.

342

343

Methods

Participants

76 healthy volunteers participated in the study after giving written informed consent (40 female, mean age 25.3 years, range 18-52 years). All participants were naïve to the purpose of the study. All participants had normal or corrected-to normal vision and reported normal hearing. The study was approved by the human research review committee of the University of Tuebingen (approval number 432 2007 BO1) and the research review committee of the University of Birmingham (approval number ERN_11-0470P).

Stimuli

The visual spatial stimulus was a Gaussian cloud of twenty bright grey dots (0.56° diameter, vertical standard deviation 1.5° , luminance 106 cd/m^2) presented on a dark grey background (luminance 62 cd/m^2 , i.e. 71% contrast). The auditory spatial cue was a burst of white noise with a 5ms on/off ramp. To create a virtual auditory spatial cue, the noise was convolved with spatially specific head-related transfer functions (HRTFs). The HRTFs were pseudo-individualized by matching participants' head width, heights, depth and circumference to the anthropometry of subjects in the CIPIC database (Algazi et al., 2001). HRTFs from the available locations in the database were interpolated to the desired locations of the auditory cue.

Experimental design and procedure

In a spatial ventriloquist paradigm, participants were presented with audiovisual spatial signals. Participants indicated the location of the sound by pressing one of 5 spatially corresponding buttons and were instructed to ignore the visual signal. Participants did not receive any feedback on their localization response. The visual signal was a cloud of 20 dots sampled from a Gaussian. The visual clouds were re-displayed with variable horizontal standard deviations (see below) every 200 ms (i.e., at a rate of 5 Hz; Figure 1A). The cloud's location mean was temporally independently resampled from five possible locations (-10° , -5° , 0° , 5° , 10°) on each trial with the inter-trial asynchrony jittered between 1.4 and 2.8 s in steps of 200 ms. In synchrony with the change in the cloud's location, the dots changed their colour and a concurrent sound was presented. The location of the sound was sampled from $\pm 5^\circ$ visual angle with respect to the mean of the visual cloud. Observers' visual uncertainty estimate was quantified in terms of the relative weight of the auditory signal on the perceived sound location. The change in the dot's colour and the emission of the sound occurred in synchrony to enhance audiovisual binding.

378

379 *Continuous sinusoidal and random walk sequences*

380 Critically, to manipulate visual noise over time, the cloud's standard deviation changed at a rate
381 of 5Hz according to i. a sinusoidal sequence, ii. a random walk sequence 1 or iii. a random walk
382 sequence 2 (Figure 2). In all sequences the horizontal standard deviation of the visual cloud
383 spanned a range from 2-18°:

384 i. *Experiment1 - Sinusoidal sequence (Sinusoid)*: A sinusoidal sequence was generated with a
385 period of 30s. During the ~65 min of the experiment, each participant completed ~ 130 cycles
386 of the sinusoidal sequence.

387 ii. *Experiment2 - Random walk sequence 1 (RW1)*: First, we generated a random walk sequence
388 of 60 s duration using a Markov chain with 76 discrete states and transition probabilities of
389 stay (1/3), change to lower (1/3) or upper (1/3) adjacent states. To ensure that the random
390 walk sequence segment starts and ends with the same value, this initial 60 s sequence
391 segment was concatenated with its temporally reversed segment resulting in a RW sequence
392 segment of 120 s duration. Each participant was presented with this 120s RW1 sequence
393 approximately 32 times during the experiment.

394 iii. *Experiment3 - Random walk sequence 2 (RW2)*: Likewise, we created a second random-walk
395 sequence of 15 s duration using a Markov chain with only 38 possible states and transition
396 probabilities similar to above. The 15 s sequence was concatenated with its temporally
397 reversed version resulting in a 30 s sequence. The smoothness of this sequence segment was
398 increased by filtering it (without phase shift) with a moving average of 250 ms. Each
399 participant was presented with this sequence segment ~130 times.

400 Generally, a session of a Sinusoid, RW1 or RW2 sequence included 1676 trials. Because of
401 experimental problems, four sessions included only 1128, 1143 or 1295 trials. Before the
402 experimental trials, participants practiced the auditory localization task in 25 unimodal auditory
403 trials, 25 audiovisual congruent trials with a single dot as visual spatial cue and 75 trials with
404 stimuli as in the main experiment.

405

406 *Experiment 4 - Sinusoidal sequence with intermittent changes in visual noise (sinusoidal jump*
407 *sequence)*

408 To dissociate the Bayesian learner from approximate exponential discounting, we designed a
409 sinusoidal sequence (period = 30 s) with intermittent increases / decreases in visual variance
410 (Figure 5). As shown in Figure 5A, we inserted increases by 8° in visual STD at three levels of
411 visual STD: 7.2°, 8.6°, 9.6° STD. Conversely, we inserted decreases by 8° in visual STD at

412 15.3°, 16.7°, 17.7° STD. We inserted jumps selectively in the period sections of high visual
413 variance to make the jumps less apparent and maximize the chances that observers treated the
414 series as a continuous sequence. As a result, the up-jumps occurred when the increases in visual
415 variance were fastest (i.e. steeper slope), while the down-jumps occurred after sections in which
416 the visual variance was relatively constant (i.e. shallow slope). We factorially combined these
417 3 (increases) x 3 (decreases) such that each sinewave cycle included exactly one sudden
418 increase and decrease in visual STD (i.e., 9 jump types). Otherwise, the experimental paradigm
419 and stimuli were identical to the continuous sinusoidal sequence described above. During the
420 ~80 min of this experiment, each participant completed ~ 154 cycles of the sinusoidal sequence
421 including 16-18 cycles for each of the 9 jump types. This sinusoidal jump sequence was
422 expected to maximize differences in adaptation rate for the Bayesian and exponential learner.
423 If participants continuously update their estimates of the visual reliability, as opposed to using
424 a change point model (Adams and MacKay, 2007, Heilbron and Meyniel, 2019), the
425 exponential learner will weight past and present uncertainty estimates throughout the entire
426 sequence according to the same exponential function. By contrast, the Bayesian learner will
427 take into account its uncertainty about the visual reliability and therefore adapt its visual
428 reliability estimate for jumps from high to low visual variance (resp. low to high visual
429 reliability, see Figure 6) more slowly than the exponential learner (see Appendix 1).

430

431 ***Subject numbers and inclusion criteria***

432 30 of the 76 subjects participated in the sinusoidal and the RW1 sequence session. Eight
433 additional subjects participated only in the RW1 sequence session. 18 additional subjects
434 participated in the RW2 sequence session. One participant completed all three continuous
435 sequences. 20 subjects participated in the sinusoidal sequence with intermittent changes in
436 visual uncertainty. In total, we collected data from 30 participants for the sinusoidal, 38
437 participants for the RW1, 19 participants for the RW2 and 20 participants for the sinusoidal
438 jump sequence. The sample sizes of 20-38 participants were based on a pilot experiment, which
439 showed individually significant effects of past visual noise on the weighting of audiovisual
440 spatial signals in 6/6 pilot participants. From these samples, we excluded participants if their
441 perceived sound location did not depend on the current visual reliability (i.e., inclusion criterion
442 $p < 0.05$ in the linear regression; please note that this inclusion criterion is orthogonal to the
443 question of whether participants' visual uncertainty estimate depends on visual signals prior to
444 the current trial). Thus, we excluded five participants of the sinusoidal and RW1 sequence and
445 two participants from the sinusoidal jump sequence. Finally, we analysed data from 25

446 participants for the sinusoidal, 33 participants for the RW1, 19 participants for the RW2 and 18
447 participants for the sinusoidal jump sequence.

448

449 *Experimental setup*

450 Audiovisual stimuli were presented using Psychtoolbox 3.09 (Brainard, 1997, Kleiner et al.,
451 2007) (www.psychtoolbox.org) running under Matlab R2010b (MathWorks) on a Windows
452 machine (Microsoft XP 2002 SP2). Auditory stimuli were presented at ~75 dB SPL using
453 headphones (Sennheiser HD 555). As visual stimuli required a large field of view, they were
454 presented on a 30" LCD display (Dell UltraSharp 3007WFP). Participants were seated at a desk
455 in front of the screen in a darkened booth, resting their head on an adjustable chin rest. The
456 viewing distance was 27.5 cm. This setup resulted in a visual field of approx. 100°. Participants
457 responded via a standard QWERTY keyboard. Participants used the buttons [i, 9, 0, -, =] with
458 their right hand for localization responses.

459

460 *Data analysis*

461 *Continuous sinusoidal and random walk sequences*

462 At trial onset the visual cloud's location mean was independently resampled from five possible
463 locations (-10°, -5°, 0°, 5°, 10°). Concurrently, the cloud's dots changed their colour and a
464 sound was presented sampled from $\pm 5^\circ$ visual angle with respect to the mean of the visual
465 cloud. The inter-trial asynchrony was jittered between 1.4 and 2.8 s in steps of 200 ms.
466 Therefore, across the experiment the trial onsets occurred at different times relative to the period
467 of the changing visual cloud's STD resulting in a greater effective sampling rate than provided
468 if the inter-trial asynchrony had been fixed.

469 For each period of the three continuous sinusoidal and random walk sequences, we
470 sorted the trials (i.e. trial-specific visual cloud's STD, visual location, auditory location and
471 observers' sound localization responses) into 20 temporally adjacent bins that covered one
472 complete period of the changing visual STD. This resulted in about 1676 trials in total/20 bins
473 = approximately 80 trials on average per bin in each subject (more specifically: a range of 52-
474 96 (Sin), 52-92 (RW 1) or 71-93 (RW2) trials, for details see Supplementary file 1-Table 1).

475 We quantified the influence of the auditory and visual locations on observers' perceived
476 auditory location for each bin by estimating a regression model separately for each bin (i.e. one
477 regression model per bin). For instance, for bin = 1 we computed:

$$478 \quad (1) \quad R_{A,\text{trial},\text{bin}=1} = L_{A,\text{trial},\text{bin}=1} \beta_{A,\text{bin}=1} + L_{V,\text{trial},\text{bin}=1} \beta_{V,\text{bin}=1} + \beta_{\text{const},\text{bin}=1} +$$

479 $e_{\text{trial},\text{bin}=1}$

480 with $R_{A,trial,bin=1}$ = Localization response for trial t and bin 1; $L_{A,trial,bin=1}$ or $L_{V,trial,bin=1}$ =
 481 ‘true’ auditory or visual location for trial t and bin 1; $\beta_{A,bin=1}$ or $\beta_{V,bin=1}$ = auditory or visual
 482 weight for bin 1; $\beta_{const,bin=1}$ = constant term; $e_{trial,bin=1}$ = error term for trial t and bin 1. For
 483 each bin b, we thus obtained one auditory and one visual weight estimate. The *relative* auditory
 484 weight for a particular bin was computed as $w_{A,bin} = \beta_{A,bin} / (\beta_{A,bin} + \beta_{V,bin})$ (Figure 2A-C).

485 By design, the temporal evolution of the physical visual variance (i.e., STD of the visual
 486 cloud) is symmetric for each period in the sinusoidal, RW1 and RW2 sequences. In other words,
 487 for physical visual noise, the 1st half and the flipped 2nd half within a period are identical (Figure
 488 3E). Given this symmetry constraint, we evaluated the influence of past visual noise on
 489 participants’ auditory weight $w_{A,bin}$ by comparing the w_A for the bins in the 1st half and the
 490 flipped 2nd half in a repeated measures ANOVA. If human observers estimate visual uncertainty
 491 by combining prior with current visual uncertainty estimates as expected for a Bayesian learner,
 492 w_A should differ between the 1st half and the mirror-symmetric flipped 2nd half of the sequence.
 493 More specifically, w_A should be smaller for the 1st half in which visual variance increased than
 494 for the mirror-symmetric time points of the 2nd half in which visual variance decreased. To test
 495 this prediction, we entered the subject-specific $w_{A,bin}$ into 2 (1st vs. flipped 2nd half) x 9 (bins,
 496 i.e. removing the bins at maximal and minimal visual noise values) repeated measures
 497 ANOVAs separately for the sinusoidal, RW1 and RW2 experiments (Table 1). For the
 498 sinusoidal sequence, we expected a main effect of ‘half’ because the sequence
 499 increased/decreased monotonically within each half period. For the RW1 and RW2 sequences,
 500 an influence of prior visual noise might also be reflected in an interaction effect of ‘half x bin’
 501 because these sequences increased/decreased non-monotonically within each half period.

502 To further test whether the noise of past visual signals influenced observers’ current
 503 visual uncertainty estimate, we employed a regression model in which the relative auditory
 504 weights $w_{A,bin}$ were predicted by the visual STD in the current bin and the difference in STD
 505 between the current and the previous bin:

$$506 \quad (2) \quad w_{A,bin} = \sigma_{V,bin} \beta_{\sigma V} + (\sigma_{V,bin} - \sigma_{V,bin-1}) \beta_{\Delta\sigma V} + \beta_{const} + e_{bin}$$

507 with $w_{A,bin}$ = relative auditory weight in bin b; $\sigma_{V,bin}$ = mean visual STD in current bin b or
 508 previous bin b-1; β_{const} = constant term; e_{bin} = error term. To allow for generalization to the
 509 population level, the parameter estimates ($\beta_{\sigma V}$, $\beta_{\Delta\sigma V}$) for each participant were entered into two-
 510 sided one-sample t-tests at the between-subject random-effects level.

511

512 *Sinusoidal sequence with intermittent changes in visual uncertainty*

513 For each period of the sinusoidal sequence with intermittent changes, we sorted the values for
514 the physical visual cloud's variance (i.e., the cloud's STD) and sound localization responses
515 into 15 temporally adjacent bins which were positioned to capture the jumps in visual noise.
516 For analysis of these sequences, we recombined the first and second halves of the 3 (increases
517 at low, middle, high) x 3 (decreases at low, middle, high) sinewave cycles into three types of
518 sinewave cycles such that both jumps were at low (= outer jump), middle (=middle jump) or
519 high (= inner jump) visual noise. This recombination makes the simplifying assumption that
520 the jump position of the first half will have negligible effects on participants' uncertainty
521 estimates of the second half. As a result of this recombination, each bin comprised at least 44-
522 51 trials across participants (Supplementary file 1-Table 1). As for the continuous sequences,
523 we quantified the auditory and visual influence on the perceived auditory location for each bin
524 based on separate regression models for the 15 temporally adjacent bins. For instance, for bin
525 = 1 we computed: $R_{A,trial,bin=1} = L_{A,trial,bin=1} * \beta_{A,bin=1} + L_{V,trial,bin=1} * \beta_{V,bin=1} + \beta_{const,bin=1} + \epsilon_{trial,bin=1}$.
526 Next, we independently computed the relative auditory weight $w_{A,bin} = \beta_{A,bin} / (\beta_{A,bin} + \beta_{V,bin})$
527 for each of the 15 temporally adjacent bins. We statistically evaluated the influence of past
528 visual noise on participants' auditory weight on the w_A in terms of the difference between 1st
529 half and flipped 2nd half using a 2 (1st vs. flipped 2nd half) x 7 (bins) x 3 (jump: inner, middle,
530 outer) repeated measures ANOVAs (Table 1).

531

532 *Computational Models (for continuous and discontinuous sequences)*

533 To further characterize whether and how human observers use their uncertainty about previous
534 visual signals to estimate their uncertainty of the current visual signal, we defined and compared
535 three models in which visual reliability (λ_V) was (1) estimated instantaneously for each trial
536 (i.e., instantaneous learner), was updated via (2) Bayesian learning or (3) exponential
537 discounting (i.e. exponential learner) (Figure 1-figure supplement 1).

538 In the following, we will first describe the generative model that accounts for the fact
539 that (1) visual uncertainty usually changes slowly across trials (i.e. time-dependent uncertainty
540 changes) and (2) auditory and visual signals can be generated by one common or two
541 independent sources (i.e. causal structure). Using this generative model as a departure point, we
542 then describe how the instantaneous learner, the Bayesian learner and the exponential learner
543 perform inference. Finally, we will explain how we account for participants' internal noise and
544 predict participants' responses from each model (i.e. the experimenter's uncertainty).

545

546 *Generative model*

547 On each trial t , the subject is presented with an auditory signal A_t , from a source $S_{A,t}$, (see Figure
 548 1-figure supplement 1) together with a visual cloud of dots at time t arising from a source, $S_{V,t}$,
 549 drawn from a Normal distribution $S_{V,t} \sim N(0, 1/\lambda_S)$ with the spatial reliability (i.e., inverse of
 550 the spatial variance): $\lambda_S = 1/\sigma_S^2$. Critically, $S_{A,t}$ and $S_{V,t}$, can either be two independent sources
 551 ($C = 2$) or one common source ($C=1$): $S_{A,t} = S_{V,t} = S_t$ (Kording et al., 2007).

552 We assume that the auditory signal is corrupted by noise, so that the internal signal is $A_t \sim$
 553 $N(S_{A,t}, 1/\lambda_A)$. By contrast, the individual visual dots (presented at high visual contrast) are
 554 assumed to be uncorrupted by noise, but presented dispersed around the location $S_{V,t}$ according
 555 to $V_{i,t} \sim N(U_t, 1/\lambda_{V,t})$, where $U_t \sim N(S_{V,t}, 1/\lambda_{V,t})$. The dispersion of the individual dots, $1/\lambda_{V,t}$,
 556 is assumed to be identical to the uncertainty about the visual mean, allowing subjects to use the
 557 dispersion as an estimate of the uncertainty about the visual mean.

558 The visual reliability of the visual cloud, $\lambda_{V,t} = 1/\sigma_{V,t}^2$, varies slowly at the re-display rate of 5
 559 Hz according to a log random walk: $\log \lambda_{V,t} \sim N(\log \lambda_{V,t-1}, 1/\kappa)$ with $1/\kappa$ being the variability
 560 of $\lambda_{V,t}$ in log space. We also use this log random walk model to approximate learning in the
 561 four jump sequence (see (Behrens et al., 2007)).

562 The generative models of the instantaneous, Bayesian and exponential learners all account for
 563 the causal uncertainty by explicitly modeling the two potential causal structures. Yet, they differ
 564 in how they estimate the visual uncertainty on each trial, which we will describe in greater detail
 565 below.

566

567 *Observer Inference*

568 The instantaneous, Bayesian and exponential learners invert this (or slightly modified, see
 569 below) generative model during perceptual inference to compute the posterior probability of
 570 the auditory location, $S_{A,t}$, given the observed A_t and $V_{i,t}$. The observer selects a response based
 571 on the posterior using a subjective utility function which we assume to be the minimization of
 572 the squared error $(S_{A,t} - S_{true})^2$. For all models, the estimate for the location of the auditory source
 573 is obtained by averaging the auditory estimates under the assumption of common and
 574 independent sources by their respective posterior probabilities (i.e. model averaging, see Figure
 575 1-figure supplement 1):

$$576 \quad (3) \quad \hat{S}_{A,t} = \hat{S}_{A,C=1,t} P(C_t = 1 | A_t, V_{1:n,t}) + \hat{S}_{A,C=2,t} (1 - P(C_t = 1 | A_t, V_{1:n,t}))$$

577 where $\hat{S}_{A,C=1,t}$ and $\hat{S}_{A,C=2,t}$ depend on the model (see below), and $P(C = 1 | A_t, V_{1:n,t})$ is the
 578 posterior probability that the audio and visual stimuli originated from the same source according
 579 to Bayesian causal inference (Kording et al., 2007).

580 (4) $P(C_t = 1|A_t, V_{1:n,t}) =$
581
$$\frac{(P(A_t, V_{1:n,t} | C = 1)P(C_t = 1))}{(P(A_t, V_{1:n,t} | C_t = 1)P(C_t = 1) + (P(A_t, V_{1:n,t} | C_t = 2)(1 - P(C_t = 1)))}$$

582 Finally, for all models we assume that the observer pushes the button associated with the
583 position closest to $\hat{S}_{A,t}$. In the following, we describe the generative and inference models for
584 the instantaneous, Bayesian and exponential learners. For the Bayesian learner, we focus
585 selectively on the model component that assumes a common cause, $C = 1$ (for full derivation
586 including both model components, see Appendix 2).

587
588 *Model 1: Instantaneous learner*

589 The instantaneous learning model ignores that the visual reliability (i.e., the inverse of visual
590 uncertainty) of the current trial depends on the reliability of the previous trial. Instead, it
591 estimates the visual reliability independently for each trial from the spread of the cloud of visual
592 dots:

593 (5) $P(S_{A,t}, U_t, \lambda_{V,t} | A_{1:t}, V_{1:n,1:t}) = P(S_{A,t}, U_t, \lambda_{V,t} | A_t, V_{1:n,t}) =$
594 $P(C = 1 | A_t, V_{1:n,t})P_{C=1}(S_t, U_t, \lambda_{V,t} | A_t, V_{1:n,t}) +$
595 $P(C = 2 | A_t, V_{1:n,t})P_{C=2}(S_{A,t}, U_t, \lambda_{V,t} | A_t, V_{1:n,t}) =$
596
$$P(C = 1 | A_t, V_{1:n,t}) \frac{P(S_t)P(A_t|S_t)P_{C=1}(U_t|S_t, \lambda_{V,t})P(V_{1:n,t}|U_t, \lambda_{V,t})P(\lambda_{V,t})}{Z_1} +$$

597 $(1 - P(C = 1 | A_t, V_{1:n,t})) P(S_{A,t})P(A_t|S_{A,t})/Z_2.$

598 with Z_1, Z_2 as normalization constants.

599 Apart from $P(C = 1 | A_t, V_t)$, these terms are all normal distributions, while we assume in this
600 model that $P(\lambda_{V,t})$ is uninformative. Hence, visual reliability is computed from the variance:

601 $\widehat{\lambda}_{V,t} = 1/(\sigma_{V,t}^2 + \frac{\sigma_{V,t}^2}{n})$ where $\sigma_{V,t}^2 = 1/(n-1) \sum_{i=1}^n (V_{i,t} - \bar{V}_{i,t})^2$ is the sample variance (and
602 $\bar{V}_t = 1/n \sum_{i=1}^n V_{i,t}$ is the sample mean). The causal component estimates are given by:

603 (6)
$$\hat{S}_{A,C=1,t} = \frac{\widehat{\lambda}_{V,t}\bar{V}_t + \lambda_A A_t}{\lambda_{V,t} + \lambda_A + \lambda_S}$$

604

605 (7)
$$\hat{S}_{A,C=2,t} = \frac{\lambda_A A_t}{\lambda_A + \lambda_S}$$

606 These two components are then combined based on the posterior probabilities of common and
607 independent cause models (see equation 3). This model is functionally equivalent to a Bayesian
608 causal inference model as described in Koerding et al. (2007), but with visual reliability

609 computed directly from the sample variance rather than a fixed unknown parameter (which the
610 experimenter estimates during model fitting).

611

612 *Model 2: Bayesian learner*

613 The Bayesian learner capitalizes on the slow changes in visual reliability across trials and
614 combines past and current inputs to provide a more reliable estimate of visual reliability and
615 hence auditory location. It computes the posterior probability based on all auditory and visual
616 signals presented until time t (here only shown for $C=1$, see Appendix 2).

617 According to Bayes rule, the joint probability of all variables until time t can be written based
618 on the generative model as:

$$619 \quad (8) \quad P(\lambda_{V,1:t}, A_{1:t}, U_{1:t}, V_{1:n,1:t}, S_{1:t}) =$$

$$620 \quad P(A_1|S_1)P(V_{1:n,1}|U_1, \lambda_{V,1})P(U_1|S_1, \lambda_{V,1})P(S_1)P(\lambda_{V,1})$$

$$621 \quad \prod_{k=2}^t P(A_k|S_k)P(V_{1:n,k}|U_k, \lambda_{V,k})P(U_k|S_k, \lambda_{V,k})P(\lambda_{V,k}|\lambda_{V,k-1})P(S_k)$$

622 As above, the visual likelihood is given by the product of individual Normal distributions for
623 each dot i : $P(V_{1:n,t}|U_t, \lambda_{V,t}) = \prod_{i=1}^n N(V_{i,t}|U_t, 1/\lambda_{V,t})$, and $P(U_t|S_t, \lambda_{V,t}) = N(U_t|S_t, 1/\lambda_{V,t})$.

624 The prior $P(S_t)$ is a Normal distribution $N(S_t|0, 1/\lambda_S)$ and the auditory likelihood

625 $P(A_t|S_t)$ is a Normal distribution $N(A_t|S_t, 1/\lambda_A)$. As described in the generative model,
626 $P(\lambda_{V,k}|\lambda_{V,k-1})$ is given by $\log \lambda_{V,t} \sim N(\log \lambda_{V,t-1}, 1/\kappa)$.

627 Importantly, only the visual reliability, $\lambda_{V,t}$, is directly dependent on the previous trial
628 ($P(\lambda_{V,k}, \lambda_{V,k-1}) = P(\lambda_{V,k}|\lambda_{V,k-1})P(\lambda_{V,k-1}) \neq P(\lambda_{V,k})P(\lambda_{V,k-1})$). Because of the Markov
629 property (i.e. $\lambda_{V,t}$ depends only on $\lambda_{V,t-1}$), the joint distribution for time t can be written as

$$630 \quad (9) \quad P(\lambda_{V,t}, \lambda_{V,t-1}, A_t, U_t, V_{1:n,t}, S_t) =$$

$$631 \quad P(A_t|S_t)P(U_t|S_t, \lambda_{V,t})P(V_{1:n}|U_t, \lambda_{V,t})P(\lambda_{V,t}|\lambda_{V,t-1})P(\lambda_{V,t-1}|V_{1:n,t-1}, A_{t-1})P(S_t)$$

$$632 \quad .$$

633 Hence, the joint posterior probability over location and visual reliability given a stream of
634 auditory and visual inputs can be rewritten as:

$$635 \quad (10) \quad P(S_t, U_t, \lambda_{V,t} | A_{1:t}, V_{1:n,1:t}) =$$

$$636 \quad P(S_t)P(A_t|S_t)P(U_t|S_t, \lambda_{V,t})P(V_{1:n}|U_t, \lambda_{V,t}) \int P(\lambda_{V,t}|\lambda_{V,t-1})P(\lambda_{V,t-1}|V_{1:n,t-1}, A_{t-1}) d\lambda_{V,t-1}/$$

$$637 \quad Z.$$

638 As this equation cannot be solved analytically, we obtain an approximate solution by factorizing
639 the posterior in terms of the unknown variables $(S_t, U_t, \lambda_{V,t})$ according to the method of
640 variational Bayes (Bishop, 2006). In this approximate method (for details see Appendix 2), the

641 posterior is factorized into three terms, each a normal distribution: $P(S_t, U_t, \lambda_{V,t_t} | A_t, V_{1:n,t}) \approx$

642 $q(S_t, U_t, \lambda_{V,t_t}) = q(S_t) * q(U_t) * q(\lambda_{V,t_t})$.

643 In order to estimate the set of parameters (mean and variance) of $q(S_t)$, $q(U_t)$ and $q(\lambda_{V,t_t})$,
644 the Free Energy is minimized iteratively (and thereby the Kullback–Leibler divergence between
645 the true and approximate distribution), until a convergence criterion is reached (here, the change
646 in each fitted parameter is less than 0.0001 between iterations).

647 This is done separately for the common cause model component (C=1) and the independent
648 cause model component (C=2). The auditory location, for the common cause model is based on
649 the approximation over the posterior location of $\hat{S}_{A,C=1,t}$ from, $q_1(S_t) = N(\hat{S}_{A,C=1,t}, \sigma_{1,t})$. The
650 auditory location for the independent cause model is simply computed as $\hat{S}_{A,C=2,t} = A_t / (1 +$
651 $\sigma_A^2 / \sigma_0^2)$, because it is independent of the visual signal.

652 The marginal model evidence is estimated based on the minimized Free Energy for each mode
653 component, $P(A_t, V_{1:n,t} | C = 1)$, respectively $P(A_t, V_{1:n,t} | C = 2)$ to form the posterior
654 probability $P(C = 1 | A_t, V_{1:n,t})$, as described above in equation 4. These values can then be
655 used to compute the predicted responses for a particular participant according to equation 3.

656

657 *Model 3: Exponential learner*

658 Finally, the observer may approximate the full Bayesian inference of the Bayesian learner by a
659 more simple heuristic strategy of exponential discounting. In the exponential discounting
660 model, the observer learns the visual reliability by exponentially discounting past visual
661 reliability estimates:

662 (11) $\hat{\lambda}_{V,t-1} = 1/\sigma_{V_t}^2 (1 - \gamma) + \hat{\lambda}_{V,t-1}\gamma$

663 where $\sigma_{V_t}^2 = 1/(n - 1) \sum_{i=1}^n (V_{i,t} - \bar{V}_t)^2$ is the sample variance and $\bar{V}_t = 1/n \sum_{i=1}^n V_{i,t}$ is the
664 sample mean.

665

666 Similar to the optimal Bayesian learner (above), this observer model uses the past to compute
667 the current reliability, but it does so based on a fixed learning rate $1 - \gamma$. Computation is
668 otherwise performed in accordance with models 1 and 2, equations 3-4 and 6-7.

669

670 *Assumptions of the computational models: motivation and caveats*

671 Computational models inherently make simplifying assumptions about the generation of the
672 sensory inputs and observers' inference.

673 First, we modelled that visual signals (i.e. the cloud's mean) were sampled from a
674 Gaussian, while they were sampled from a uniform discrete distribution (i.e. $[-10^\circ, -5^\circ, 0^\circ, 5^\circ,$
675 $10^\circ]$) in the experiment. Gaussian assumptions about the stimuli locations have nearly
676 exclusively been made in the recent series of studies focusing on Bayesian Causal Inference in
677 multisensory perception (Kording et al., 2007, Rohe and Noppeney, 2015b, Rohe and
678 Noppeney, 2015a). Because visual signals have been sampled from a wide range of visual angle
679 (i.e. 20°) and are corrupted by physical (i.e. cloud of dots) and internal neural noise, we used
680 the simplifying assumption of a Gaussian spatial prior consistent with previous research.

681 Second, we assumed that the auditory signal location is sampled from a Gaussian, while
682 the experiments presented sounds $\pm 5^\circ$ from the visual location. These Gaussian assumptions
683 about sound location can be justified by the fact that observers are known to be limited in their
684 sound localization ability, particularly when generic head related transfer functions were used
685 to generate spatial sounds. Moreover, because sounds are presented together with visual signals,
686 it is even harder for observers to obtain an accurate estimate of the sound's location.

687 Third, in the experiment we generated the cloud of dots directly from a Gaussian
688 distribution centred on S_t . By contrast, in the model we introduced a hidden variable U_t that is
689 sampled from a Gaussian centred on S_t . The visual cloud of dots is then centred on this hidden
690 variable U_t . We introduced this additional hidden variable U_t to account for observers'
691 additional causal uncertainty in natural environments, in which even signals from a common
692 source may not fully coincide in space. Critically, the dispersion of the cloud of dots is set to
693 be equal to the standard deviation of the distribution from which U_t is sampled, so that the
694 cloud's standard deviation informs observers about the variance of the hidden variable U_t .

695

696 *Inference by the experimenter*

697 From the observer's viewpoint, this completes the inference process. However, from the
698 experimenter's viewpoint, the internal variable for the auditory stimulus, A_t , is unknown and
699 not directly under the experimenter's control. To integrate out this unknown variable, we

700 generated 1,000 samples of the internal auditory value for each trial from the generative process
701 $A_t \sim N(S_{A,t,true}, \sigma_A^2)$, where $S_{A,t,true}$ was the true location the auditory stimulus came from. For
702 each value of A_t , we obtained a single estimate $\hat{S}_{A,t}$ (as described above). To link these estimates
703 with observers' button response data, we assumed that subjects push the button associated with
704 the position closest to $\hat{S}_{A,t}$. In this way, we obtained a histogram of responses for each subject
705 and trial which provide the likelihood of the model parameters given a subject's responses:
706 $P(resp_t | \kappa, \sigma_A, P_{common}, S_{A,t,true}, S_{V,t,true})$.

707

708 ***Model estimation and comparison***

709 Parameters for each model (for all models: σ_A , $P_{common} = P(C=1)$, σ_0 , Bayesian learner: κ ,
710 exponential learner: γ) were fit for each individual subject by sampling using a symmetric
711 proposal Metropolis-Hasting (MH) algorithm (with A_t integrated out via sampling, see above).
712 The MH algorithm iteratively draws samples set_n from a probability distribution through a
713 variant of rejection sampling: if the likelihood of the parameter set is larger than the previous
714 set, the new set is accepted, otherwise it is accepted with probability $L(model|set_n)/L(model|set_{n-1})$,
715 where $L(resp|set_n) = \prod_t P(resp_t | \kappa, \sigma_A, P_{common}, S_{A,t,true}, S_{V,t,true})$ (for Bayesian learner).
716 We sampled 4000 steps from 4 sampling chains with thinning (only using every 4th sampling
717 to avoid correlations in samples), giving a total of 4000 samples per subject data sets.
718 Convergence was assessed through scale reduction (using criterion $R < 1.1$ (Gelman et al.,
719 2013)). Using sampling does not just provide a single parameter estimate for a data set (as when
720 fitting maximum likelihood), but can instead be used to assess the uncertainty in estimation for
721 the data set. The model code was implemented in Matlab (Mathworks, MA) and ran on two
722 dual Xeon workstations. Each sample step, per subject data set, took 30 seconds on a single
723 core (~42 hours per sampling chain).

724 Quantitative Bayesian model comparison of the three candidate models was based on
725 the Watanabe-Akaike Information Criterion (WAIC) as an approximation to the out of sample
726 expectation (Gelman et al., 2013). At the fixed-effects level, Bayesian model comparison was
727 performed by summing the WAIC over all participants within each experiment. For a random-
728 effects analysis, we transformed the WAIC into log-likelihoods by dividing them by minus 2.
729 We then computed the protected exceedance probability that one model is better than the other
730 model beyond chance using hierarchical Bayesian model selection (Penny et al., 2010, Rigoux
731 et al., 2014).

732 To qualitatively compare the localization responses given by the participants and the
733 responses predicted by the instantaneous, Bayesian and exponential learner, we computed the

734 auditory weight w_A from the predicted responses of the three models exactly as in the analysis
735 for the behavioral data. For illustration, we show and compare the model's w_A from the 1st and
736 the flipped 2nd half of the periods for each of the four experiments (cf. Figure 3, Figure 4, Figure
737 5B/C and Figure 5-figure supplement 1).

738

739 ***Parameter recovery***

740 To test the validity of the models, we performed parameter recovery and were able to
741 recover the generating values with a bias of all parameters smaller than 10 percent (for full
742 details of bias and variance across parameters, see Appendix 1 and Supplementary file 1-Table
743 7).

744

745 ***Simulated localization responses***

746 To further compare the Bayesian and exponential learner and assess whether they can
747 be discriminated experimentally, we simulated the choices of 12 subjects for the continuous
748 sinusoidal and sinusoidal jump sequence using the Bayesian learner model (parameters: $\sigma_A = 6$
749 deg, $\kappa = 15$, $P_{\text{common}} = 0.7$ and $\sigma_0 = 12$ degrees). To increase observers' uncertainty about their
750 visual reliability estimates, we reduced the number of dots in the visual clouds from 20 to 5
751 dots where we ensured that the mean and variance of the 5 dots corresponded to the
752 experimentally defined visual mean and variance. We then fitted the Bayesian learner and
753 exponential learner models to each simulated data set (using the BADS toolbox for likelihood
754 maximization (Acerbi and Ma, 2017)). The fitted parameters for the Bayesian model, set_{Bayes}
755 were very close to the parameters used to generate observers' simulated responses (sinusoidal
756 sequence, fitted parameters: $\sigma_A = 6.11^\circ$, $\kappa = 17.5$, $P_{\text{common}} = 0.72$ and $\sigma_0 = 12.4^\circ$; sinusoidal
757 jump sequence, fitted parameters: $\sigma_A = 6.08^\circ$, $\kappa = 17.3$, $P_{\text{common}} = 0.71$ and $\sigma_0 = 12.2^\circ$) – thereby
758 providing a simple version of parameter recovery. The parameters of the exponential model,
759 set_{Exp} (fitted to observers' responses generated from the Bayesian model) were very similar to
760 those of the Bayesian learner (sinusoidal sequence: $\sigma_A = 5.99^\circ$, $\gamma = 0.70$, $P_{\text{common}} = 0.61$ and σ_0
761 $= 12.0^\circ$, sinusoidal jump sequence: $\sigma_A = 6.06^\circ$, $\gamma = 0.70$, $P_{\text{common}} = 0.65$ and $\sigma_0 = 12.0^\circ$). Moreover,
762 the fits to the simulated observers' responses were very close for the two models (Figure 6),
763 with mean log likelihood difference ($\log(L(\text{resp}/set_{\text{Bayes}})) - \log(L(\text{resp}/set_{\text{Exp}}))$) = 1.82 for the
764 sinusoidal and 2.74 for the sinusoidal jump sequence (implying a slightly better fit for the
765 Bayesian learner). Figure 6C and D show the timecourses of observers' visual uncertainty
766 (STD) as estimated by the Bayesian and exponential learners.

767

768 ***Data availability***

769 The behavioral data and model predictions as well as the code for modelling and analyses
770 scripts are available in an OSF repository: <https://osf.io/gt4jb/>

771

- 773 ACERBI, L., DOKKA, K., ANGELAKI, D. E. & MA, W. J. 2018. Bayesian comparison of explicit
774 and implicit causal inference strategies in multisensory heading perception. *PLoS*
775 *computational biology*, 14, e1006110.
- 776 ACERBI, L. & MA, W. J. 2017. Practical Bayesian optimization for model fitting with Bayesian
777 adaptive direct search. *Advances in neural information processing systems (NIPS)*, 1836-1846.
- 778 ACERBI, L., VIJAYAKUMAR, S. & WOLPERT, D. M. 2014. On the origins of suboptimality in
779 human probabilistic inference. *PLoS computational biology*, 10, e1003661.
- 780 ADAMS, R. P. & MACKAY, D. J. 2007. Bayesian online changepoint detection. *arXiv preprint*
781 *arXiv:0710.3742*.
- 782 ALAIS, D. & BURR, D. 2004. The ventriloquist effect results from near-optimal bimodal integration.
783 *Curr Biol*, 14, 257-62.
- 784 ALGAZI, V. R., DUDA, R. O., THOMPSON, D. M. & AVENDANO, C. The cipc hrtf database.
785 Applications of Signal Processing to Audio and Acoustics, 2001 IEEE Workshop on the,
786 2001. IEEE, 99-102.
- 787 ALLER, M. & NOPPENY, U. 2019. To integrate or not to integrate: Temporal dynamics of
788 Bayesian Causal Inference. *PLoS Biol*, in press.
- 789 BATTAGLIA, P. W., JACOBS, R. A. & ASLIN, R. N. 2003. Bayesian integration of visual and
790 auditory signals for spatial localization. *J Opt Soc Am A Opt Image Sci Vis*, 20, 1391-7.
- 791 BECK, J. M., MA, W. J., KIANI, R., HANKS, T., CHURCHLAND, A. K., ROITMAN, J.,
792 SHADLEN, M. N., LATHAM, P. E. & POUGET, A. 2008. Probabilistic population codes for
793 Bayesian decision making. *Neuron*, 60, 1142-1152.
- 794 BEHRENS, T. E., WOOLRICH, M. W., WALTON, M. E. & RUSHWORTH, M. F. 2007. Learning
795 the value of information in an uncertain world. *Nature neuroscience*, 10, 1214.
- 796 BERNIKER, M., VOSS, M. & KORDING, K. 2010. Learning priors for Bayesian computations in the
797 nervous system. *PLoS One*, 5.
- 798 BISHOP, C. M. 2006. *Pattern recognition and machine learning*, New York, Springer.
- 799 BRAINARD, D. H. 1997. The psychophysics toolbox. *Spatial vision*, 10, 433-436.
- 800 DRUGOWITSCH, J., DEANGELIS, G. C., KLIER, E. M., ANGELAKI, D. E. & POUGET, A. 2014.
801 Optimal multisensory decision-making in a reaction-time task. *Elife*, 3, e03005.
- 802 DRUGOWITSCH, J., WYART, V., DEVAUCHELLE, A.-D. & KOECHLIN, E. 2016. Computational
803 precision of mental inference as critical source of human choice suboptimality. *Neuron*, 92,
804 1398-1411.
- 805 ERNST, M. O. & BANKS, M. S. 2002. Humans integrate visual and haptic information in a
806 statistically optimal fashion. *Nature*, 415, 429-33.
- 807 FISER, J., BERKES, P., ORBAN, G. & LENGYEL, M. 2010. Statistically optimal perception and
808 learning: from behavior to neural representations. *Trends Cogn Sci*, 14, 119-30.
- 809 GELMAN, A., HWANG, J. & VEHTARI, A. 2014. Understanding predictive information criteria for
810 Bayesian models. *Statistics and computing*, 24, 997-1016.
- 811 GELMAN, A., STERN, H. S., CARLIN, J. B., DUNSON, D. B., VEHTARI, A. & RUBIN, D. B.
812 2013. *Bayesian data analysis*, Chapman and Hall/CRC.
- 813 HEILBRON, M. & MEYNIEL, F. 2019. Confidence resets reveal hierarchical adaptive learning in
814 humans. *PLoS computational biology*, 15, e1006972.
- 815 HOU, H., ZHENG, Q., ZHAO, Y., POUGET, A. & GU, Y. 2019. Neural Correlates of Optimal
816 Multisensory Decision Making under Time-Varying Reliabilities with an Invariant Linear
817 Probabilistic Population Code. *Neuron*.
- 818 JACOBS, R. A. 1999. Optimal integration of texture and motion cues to depth. *Vision Res*, 39, 3621-9.
- 819 JACOBS, R. A. & FINE, I. 1999. Experience-dependent integration of texture and motion cues to
820 depth. *Vision Research*, 39, 4062-4075.
- 821 KLEINER, M., BRAINARD, D., PELLI, D., INGLING, A., MURRAY, R. & BROUSSARD, C.
822 2007. What's new in Psychtoolbox-3. *Perception*, 36, 1.1-16.
- 823 KNILL, D. C. & POUGET, A. 2004. The Bayesian brain: the role of uncertainty in neural coding and
824 computation. *Trends Neurosci*, 27, 712-9.
- 825 KNILL, D. C. & RICHARDS, W. 1996. *Perception as Bayesian inference*, Cambridge University
826 Press.

827 KORDING, K. P., BEIERHOLM, U., MA, W. J., QUARTZ, S., TENENBAUM, J. B. & SHAMS, L.
828 2007. Causal inference in multisensory perception. *PLoS One*, 2, e943.
829 MA, W. J., BECK, J. M., LATHAM, P. E. & POUGET, A. 2006. Bayesian inference with
830 probabilistic population codes. *Nat Neurosci*, 9, 1432-8.
831 MA, W. J. & JAZAYERI, M. 2014. Neural coding of uncertainty and probability. *Annual review of*
832 *neuroscience*, 37, 205-220.
833 MEIJER, D., VESELIĆ, S., CALAFIORE, C. & NOPPENNEY, U. 2019. Integration of audiovisual
834 spatial signals is not consistent with maximum likelihood estimation. *Cortex*, 119, 74-88.
835 MIKULA, L., GAVEAU, V., PISELLA, L., KHAN, A. Z. & BLOHM, G. 2018. Learned rather than
836 online relative weighting of visual-proprioceptive sensory cues. *Journal of neurophysiology*,
837 119, 1981-1992.
838 NORTON, E. H., ACERBI, L., MA, W. J. & LANDY, M. S. 2019. Human online adaptation to
839 changes in prior probability. *PLoS computational biology*, 15, e1006681.
840 PENNY, W. D., STEPHAN, K. E., DAUNIZEAU, J., ROSA, M. J., FRISTON, K. J., SCHOFIELD,
841 T. M. & LEFF, A. P. 2010. Comparing families of dynamic causal models. *PLoS Comput*
842 *Biol*, 6, e1000709.
843 RIGOUX, L., STEPHAN, K. E., FRISTON, K. J. & DAUNIZEAU, J. 2014. Bayesian model selection
844 for group studies—revisited. *Neuroimage*, 84, 971-985.
845 ROHE, T., EHLIS, A.-C. & NOPPENNEY, U. 2019. The neural dynamics of hierarchical Bayesian
846 causal inference in multisensory perception. *Nature Communications*, 10, 1907.
847 ROHE, T. & NOPPENNEY, U. 2015a. Cortical hierarchies perform Bayesian causal inference in
848 multisensory perception. *PLoS Biol*, 13, e1002073.
849 ROHE, T. & NOPPENNEY, U. 2015b. Sensory reliability shapes perceptual inference via two
850 mechanisms. *Journal of Vision*, 15, 1-16.
851 ROHE, T. & NOPPENNEY, U. 2016. Distinct computational principles govern multisensory integration
852 in primary sensory and association cortices. *Current Biology*, 26, 509-514.
853 SHEN, S. & MA, W. J. 2016. A detailed comparison of optimality and simplicity in perceptual
854 decision making. *Psychological review*, 123, 452.
855 TRIESCH, J., BALLARD, D. H. & JACOBS, R. A. 2002. Fast temporal dynamics of visual cue
856 integration. *Perception*, 31, 421-434.
857 VAN BEERS, R. J., SITTING, A. C. & GON, J. J. D. V. D. 1999. Integration of proprioceptive and
858 visual position-information: An experimentally supported model. *Journal of neurophysiology*,
859 81, 1355-1364.
860 WOZNY, D. R., BEIERHOLM, U. R. & SHAMS, L. 2010. Probability matching as a computational
861 strategy used in perception. *PLoS Comput Biol*, 6.
862 ZEMEL, R. S., DAYAN, P. & POUGET, A. 1998. Probabilistic interpretation of population codes.
863 *Neural computation*, 10, 403-430.

864

865

Legends of supplementary figures

866 **Figure 1-figure supplement 1. Generative model for the Bayesian learner.** The Bayesian Causal
867 Inference model explicitly models whether auditory and visual signals are generated by one common
868 ($C=1$) or two independent sources ($C=2$) (for further details see Koerding et al., 2007). We extend this
869 Bayesian Causal Inference model into a Bayesian learning model by making the visual reliability ($\lambda_{V,t}$,
870 i.e. the inverse of uncertainty or variance) of the current trial dependent on the previous trial.

871

872 **Figure 2-figure supplement 1. Time course of the relative auditory weights for continuous**
873 **sequences of visual noise when controlling for location of the cloud of dots in the previous trial.**

874 Relative auditory weights (mean across participants \pm SEM, left ordinate) and visual noise (i.e., STD of
875 the cloud of dots, right ordinate) are displayed as a function of time as shown in Figure 2 of the main
876 text. To compute the relative auditory weights, the sound localization responses were regressed on the
877 A and V signal locations within bins of 1.5 s (**A, B**) or 6 s (**C**) width across sequence repetitions within
878 each participant. To control for a potential effect of past visual locations, the location of the visual cloud
879 of dots in the previous trial was included in this regression model as a covariate (cf. Supplementary file
880 1-Table 3).

881

882 **Figure 5-figure supplement 1. Time course of relative auditory weights and visual noise for the**
883 **sinusoidal sequence with intermittent jumps in visual noise.**

884 Relative auditory weights $w_{A,\text{bin}}$ (mean
885 across participants) of the 1st (solid) and the flipped 2nd half (dashed) of a period (binned into 15 time
886 bins) plotted as a function of the time in the sinusoidal sequence with intermitted inner (light gray),
887 middle (gray) and outer (dark gray) jumps. Relative auditory weights were computed from auditory
888 localization responses of exponential (**A**) or instantaneous (**B**) learning models. For comparison, the
889 standard deviation of the visual signal is shown in (**C**). Please note that all models were fitted to
890 observers' auditory localization responses (i.e. not the auditory weight w_A).

890

891 **Figure 5-figure supplement 2. Time course of relative auditory weights and root mean squared**
892 **error of the computational models before and after the jumps in the sinusoidal sequence with**

893 **intermittent jumps. (A)** Relative auditory weights w_A (mean across participants) shown as a function
894 of time around the up-jumps (left panel) and the down-jumps (right panel) for observers' behavior, the
895 instantaneous, exponential and Bayesian learner. Relative auditory weights were computed from
896 auditory localization responses for behavioral data and for the predictions of the three computational
897 models in time bins of 200 ms (i.e., 5Hz rate of the visual clouds). Trials from the three types of up- and
898 down-jumps were pooled to increase the reliability of the w_A estimates. Because time bins included only
899 few trials in some participants, individual w_A values that were smaller or larger than the three times the
900 scaled median absolute deviation were excluded from the analysis. Note that the up jumps occurred

901 around the steepest increase in visual noise, so that the Bayesian and exponential learners
902 underestimated visual noise (cf. Figure 5C), leading to smaller w_A as compared to the instantaneous
903 learner already before the up jump. **(B)** Root mean squared error (RMSE; computed across participants)
904 between w_A computed from behavior and the models' predictions (as shown in A), shown as a function
905 of the time around the up-jumps (left panel) and the down-jumps (right panel). Please note that all
906 models were fitted to observers' auditory localization responses (i.e. not the auditory weight w_A).

907
908
909

Legends of appendix figures

910 **Appendix 2-figure 1.** Generative model, for one ($C=1$) or two sources ($C=2$).

911

912 **Appendix 2-figure 2.** Approximation of theta using Laplace approximation.

913

914 **Appendix 2-figure 3.** Comparing variational Bayes approximation with a numerical discretised grid
915 approximation. Top row: Example visual stimuli over eight subsequent trials. Middle row: The
916 distribution of estimated sample variance, with no learning over trials. Bottom row: The distribution of
917 \hat{V}_t for the Bayesian model that incorporates the learning across trials. Red line is the numerical
918 comparison when using a discretised grid to estimate variance, as opposed to the variational Bayes
919 (green line).

920 **Acknowledgements**

921 This study was funded by the ERC (ERC-multisens, 309349), the Max Planck Society and the
922 Deutsche Forschungsgemeinschaft (DFG; grant number RO 5587/1-1).

923

924 **Competing interests**

925 The authors report no conflict of interest.