# Evidence integration and decision confidence are modulated by stimulus consistency

Moshe Glickman[1,2,✉], Rani Moran[2,3,6] and Marius Usher[4,5,6,✉]

[1]Department of Experimental Psychology, University College London, London, UK. [2]Max Planck UCL Centre for Computational Psychiatry and Ageing Research, University College London, London, UK. [3]Wellcome Centre for Human Neuroimaging, University College London, London, UK. [4]School of Psychology, University of Tel Aviv, Tel Aviv, Israel. [5]Sagol School of Neuroscience, University of Tel Aviv, Tel Aviv, Israel. [6]These authors jointly supervised this work: Rani Moran, Marius Usher. ✉Corresponding authors.

Moshe Glickman: mosheglickman345@gmail.com

Marius Usher: marius@tauex.tau.ac.il

Evidence integration is a normative algorithm for choosing between alternatives with noisy evidence, which has been successful in accounting for vast amounts of behavioural and neural data. However, this mechanism has been challenged by non-integration heuristics, and tracking decision boundaries has proven elusive. Here we first show that the decision boundaries can be extracted using a model-free behavioural method termed decision classification boundary, which optimizes choice classification based on the accumulated evidence. Using this method, we provide direct support for evidence integration over non-integration heuristics, show that the decision boundaries collapse across time and identify an integration bias whereby incoming evidence is modulated based on its consistency with preceding information. This consistency bias, which is a form of pre-decision confirmation bias, was supported in four cross-domain experiments, showing that choice accuracy and decision confidence are modulated by stimulus consistency. Strikingly, despite its seeming sub-optimality, the consistency bias fosters performance by enhancing robustness to integration noise.

Decisions often require the integration of multiple, potentially contradictory, pieces of evidence. Consider, for example, a judge deliberating over whether a defendant is guilty or not, or a medical doctor diagnosing a patient's disease. Extensive research has converged on the proposition that integration of evidence to a decision boundary is a normative mechanism for such evidence-based decisions. This mechanism provides the fastest mean response time (RT) for a target accuracy rate[1–4] and accounts for an impressive amount of behavioural and neural choice data (see ref.[5] for a review). For instance, integration-to-boundary models[1,6–11] provide a parsimonious account for the shape of choice-RT distributions of correct and incorrect responses as a function of stimulus difficulty, as well as for the well-known speed accuracy trade-off[12] stating that people can improve their decision accuracy by sampling more information, and vice versa. Moreover, integration-to-boundary models are supported by the monitoring of neural activation in brain decision areas during choice tasks[13,14] (but see ref.[15]).

Despite this strong support, the evidence integration framework has been challenged by alternative non-integration mechanisms, such as heuristics based on the detection of a single high-value sample, which can account for many of these choice-patterns as well[16] (see also discussion in ref. [17]). Moreover, research within the evidence-integration framework has suggested that, in many decision-environments, reward-rate is optimized when the choice threshold varies (for example, collapses) as a function of time[18,19]. Evidence for such time-varying boundaries have been found in both humans and non-human primates[20,21] (but see ref. [22]). These studies, however, have estimated the boundary whilst assuming arbitrary functional assumptions (for example, that the boundary decays according to a Weibull function). Thus, it is important to validate the integration assumption, and to monitor the decision boundary without imposing such assumptions.

A second aspect of the normative model is that the evidence (construed as the increase in the log likelihood of the two hypotheses) is integrated without biases or distortions. However, recent findings indicate a variety of biases in the sampling and weighting of evidence. Attentional biases, for example, affect the relative weighting assigned to simultaneous pieces of evidence[23–26]. Another type of decision biases are history biases, such as the confirmation bias[27–31], according to which committing to a categorical choice distorts the interpretation of subsequent information. Potential biases of this kind need to

be taken into account when assessing whether evidence is accumulated and when estimating decision boundaries. In turn, the decision bound can interact with the estimation of the bias itself.

Here, we addressed these issues by using a novel behavioural method, which is agnostic to the (temporally) functional shape of the boundary and which we term decision classification boundary (DCB). The DCB extracts the decision-boundary (at each time frame) by optimizing the classification of the agent's behaviour (that is, terminating the trial by choosing alternative A, terminating the trial by choosing alternative B or sampling more evidence), based on the evidence accumulated to this time point. When evidence integration is perfect (that is, free of distortions and biases), the DCB recovers the decision boundary. More broadly, however, DCBs provide a novel behavioural signature — a benchmark for evaluating biases in evidence-integration — and also allow us to derive a simplified behaviourally approximate bias-free model, which compensates for integration biases, via a change in the classification curve only (see Results section for further details and an example). Applying this method to data from experiments across choice domains (numerical cognition and perception), we find strong support for evidence integration over heuristic non-integration models. Furthermore, we demonstrate an important new factor modulating evidence accumulation, viz. stimulus consistency, corresponding to an increased relative weighting of pieces of evidence preceded by information supporting the same choice alternative, resulting in a type of momentary confirmation-bias[28,30,31], which operates during (rather than after) a decision. Importantly, this mechanism contributes to decision performance, by increasing the robustness to late (non-encoding) noise[32,33].

We start with a description of our experimental design (showing it is possible to extract behavioural signatures of integration to boundary), followed by a computational section that presents the DCB method. We then apply this method to the data from two cross-domain experiments, focusing on the dependency of the DCB on stimulus consistency, and resort to computational modeling to specify the stimulus-consistency mechanism. We then present the results of two additional experiments designed to validate the predictions of this mechanism. Finally, we examine if and under which conditions the stimulus-consistency mechanism can foster performance**.**

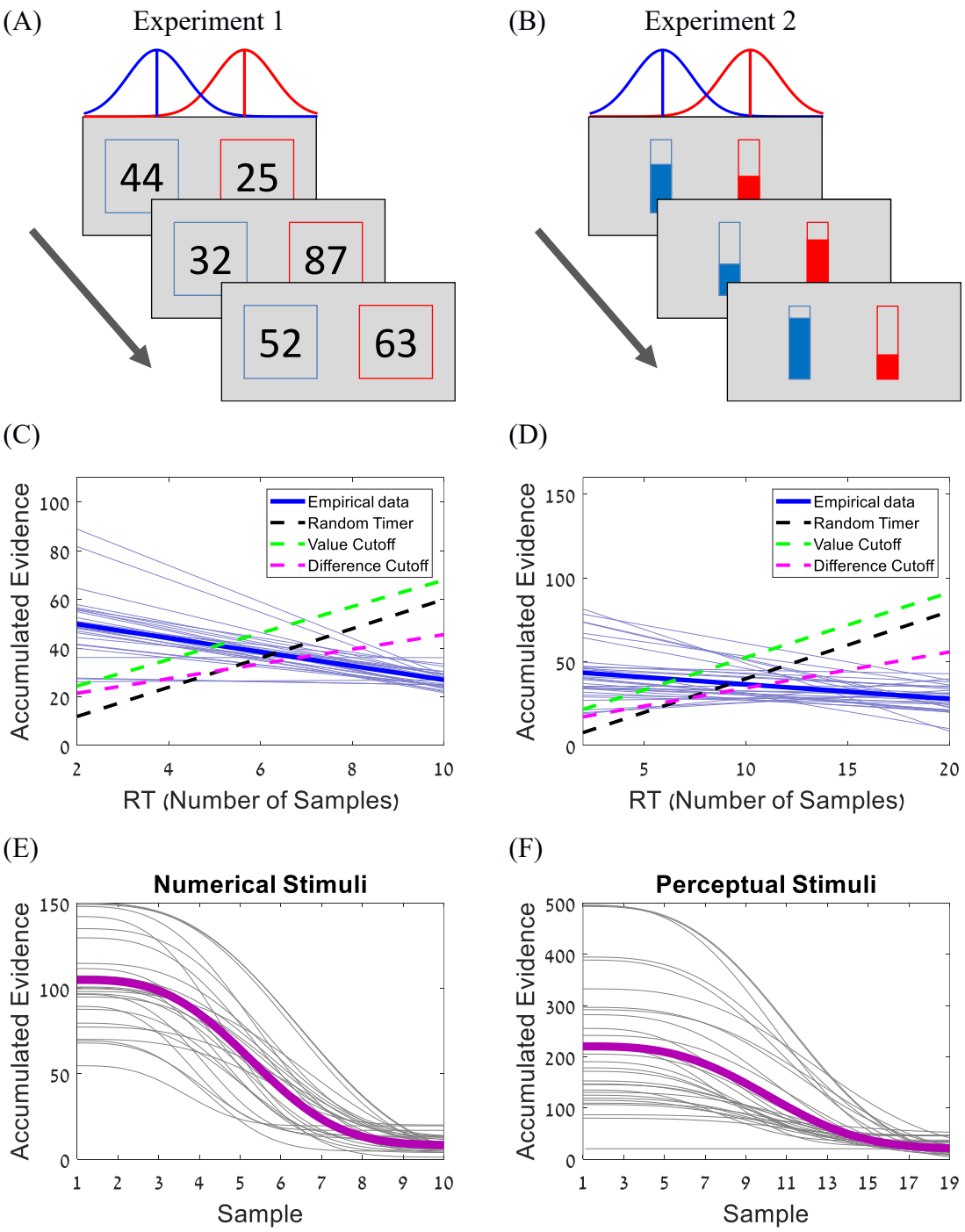

(A)
Experiment 1
44
25
32
87
52
63
(B)
Experiment 2
(C)
Empirical data
Random Timer
Value Cutoff
Difference Cutoff
Accumulated Evidence
RT (Number of Samples)
(D)
Empirical data
Random Timer
Value Cutoff
Difference Cutoff
Accumulated Evidence
RT (Number of Samples)
(E)
Numerical Stimuli
Accumulated Evidence
Sample
(F)
Perceptual Stimuli
Accumulated Evidence
Sample

***Figure. 1.*** *Experimental paradigms and behavioural signature of integration to boundary. (A-B) Experimental paradigms. Participants are presented with pairs of numerical values (A, experiment 1) or bars (B, experiment 2) sampled from two overlapping Normal distributions, and are asked to choose which sequence was drawn from a distribution with a higher mean value (experiment 1) or greater mean length (experiment 2). The presentation is terminated by the decision of the participant (that is, free response protocol). (C) The accumulated evidence of all participants (thick blue line) as a function of decision time. Thin blue lines correspond to accumulated evidence of individual participants. Dashed black, green and magenta lines correspond to the random time model, and to the value and difference heuristics, respectively. (D) As in (C) but for experiment 2. (E-F) The collapsing boundaries obtained in experiment 1 (E) and 2 (F). Thick purple lines correspond to the boundaries generated using the group mean parameters, grey lines correspond to the boundaries of individual participants.*

## Results

**Experimental design and signature of integration to boundary.** In two experiments (experiment 1, reported in ref. [34] and experiment 2, novel data), participants were presented with sequences of pairs of stimuli (two digit numbers or bars, respectively; Figure 1A-B) sampled from two overlapping Normal distributions (see Experimental methods). The sequences were presented at a rate of 2 pairs per second (numbers) and 5 pairs per second (bars), and were terminated by the participant's response. The task was to select the sequence that corresponded to the generating distribution with the higher mean (red Gaussian in Figure 1A-B). In experiment 2, participants also reported their degree of confidence for each choice. Between trials, we manipulated stimulus difficulty by varying the separation between the Normal distributions (see Experimental methods for further details). In total, 27 participants performed 500 trials in experiment 1 and 30 participants performed 480 trials in experiment 2.

In previous work, we showed that the choices in experiment 1 (numerical evidence) support integration of evidence to a collapsing boundary and excluded a set of non-integration models[34]. Here, we extend this analysis to the data in the perceptual domain (experiment 2). The thick blue lines (group data) and thin grey lines (individual subjects) in Fig. 1C-D are obtained by integrating the trial-by-trial stimulus evidence (that is, computing the cumulative sum of differences between the sequence pairs of samples) until the decision moment, and averaging across trials for each RT. These lines show a mildly decreasing pattern (Figure 1C: $b$ = -2.86, $t_{\text{against 0}}$ = -7.08, $p < 0.001$, 95% CI -3.65 to -2.07 and Figure

1D: $b$ = -0.86, $t_{\text{against 0}}$ = -3.18, $p$ = 0.001, 95% CI -1.40 to -0.33), which is the behavioural signature of integration to a collapsing boundary (for further details, see ref. [34] and Supplementary Figure 1). These results are consistent with previous findings that accumulated evidence decreases with RT[35–37], as well as with model fitting results using a Weibull parametrization of the boundary[38], which also indicates a collapsing boundary (Figure 1E-F). Critically, the descending slopes of integrated evidence rule out non-integration strategies, such as: (1) random-timer, in which the response time is determined by a process that is exogenous to the integration of evidence (whose prediction is indicated by the dashed black line), (2) value-cutoff, in which observers choose the sequence in which a number exceeding some predetermined threshold first appears (whose prediction is indicated by the dashed green line), and (3) difference-cutoff, in which observers choose based on the first frame in which the difference between the numbers exceeds a predetermined threshold (whose prediction is indicated by the dashed magenta line). All of these non-integration models predict that the integrated evidence increases (rather than decreases) with the number of samples (see Computational methods for further details about these strategies). This is because, if the stopping rule is independent of the integrated evidence, longer decision trials necessarily accumulate more evidence. Whilst these results provide support for integration to boundary, the actual shape of the boundary (Figure 1E-F) is only extracted via model fitting (note that the actual boundary are not linearly decreasing with time), as the evidence-integration lines (Figure 1C-D, blue and thin grey lines) are systematically biased by the accumulation of noise during the trial[34]. In the following, we present a model-free method to estimate the decision boundary via a classification boundary curve, which is more robust to accumulation noise and reconstructs the actual shape of the decision boundary.

**Model-Free extraction of decision boundaries**. We simulated synthetic data based on the experimental task we used in our experiments (Figure 1A-B and Computational methods), in which sequences of values are sampled from two overlapping Normal distributions. It is assumed that the subjects integrate a noisy version of the evidence at each frame, according to the following difference equation:

$$X(t) = X(t-1) + \mu(t) + \varepsilon(t),\ \varepsilon \sim N(0, \sigma^2) \qquad \text{Eq. (1)}$$

where *X(t)* is the accumulated differences between the sequences at time *t*, $\mu(t)$ is the difference between the samples at time *t* and $\varepsilon(t)$ is a temporally independent random internal Gaussian noise, which is independent from the evidence-sampling noise (Figure 1A-B). Note that, unlike $\varepsilon(t)$, $\mu(t)$ is directly available to the experimenter. We also assumed that a response is triggered when the integrated noisy evidence reaches one of two symmetric decision boundaries. Two types of boundaries were used in these simulations: fixed and collapsing[38]. For the collapsing boundary, we follow ref. [38] and parameterize the boundary using Weibull functions (see Computational methods). In each simulation, we ran 10,000 trials in which values sampled from the two distributions were integrated with additional internal noise. For each trial, we recorded the choice, as well as the input sequences (without the internal noise) sampled until a decision was made.

Based on this data, we reconstruct the boundary at each time frame using a method based on linear discriminant analysis[39,40] (LDA), which generates boundary classification curves. These curves are obtained by applying LDA to the integrated evidence excluding the random internal noise, defined as:

$$Y(t) = Y(t-1) + \mu(t) \qquad \text{Eq. (2)}$$

The LDA was applied to the integrated-evidence (*Y(t)*) of the observer at each time frame, so as to classify the action at each time frame to one of three categories: choose alternative A (Figure 2A-B, blue distributions), choose alternative B (Figure 2A-B, green distributions) or continue sampling (Figure 2A-B, pink distributions). The classification boundary curve (Figure 2A-B, red line) best separates the different classes (Computational methods section). Note that, for each time frame (*t*), the LDA was applied to all trials in the experiment that were not terminated before the *t*-th frame. In addition, note that in the absence of internal noise, the true decision boundary would separate these categories perfectly. However, in the presence of internal noise, which may increase the values of integrated evidence that lie below the boundary (blue areas), or vice versa (pink areas above the boundary), there is some unavoidable overlap.

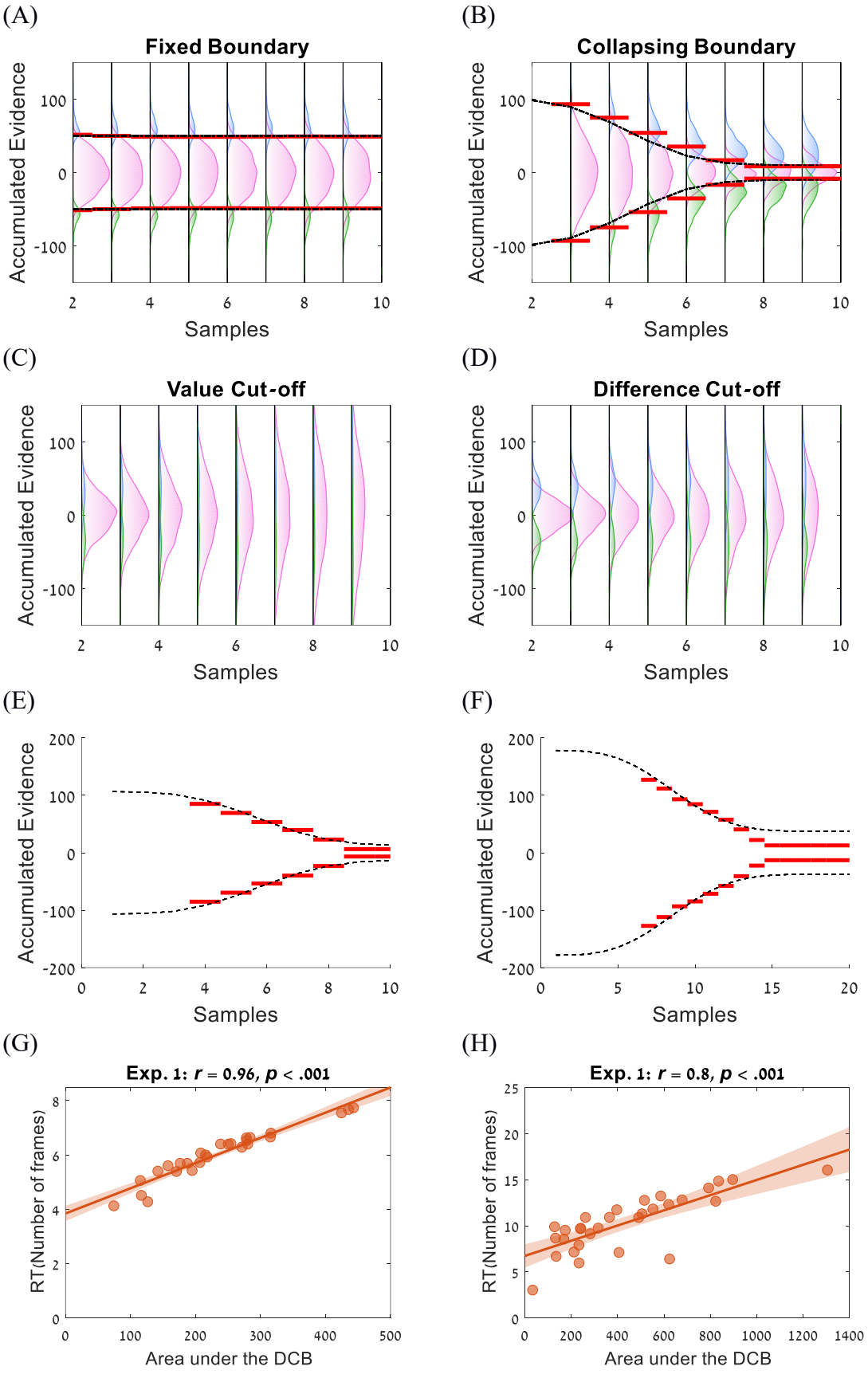

(A)
Fixed Boundary
Accumulated Evidence
Samples
(B)
Collapsing Boundary
Accumulated Evidence
Samples
(C)
Value Cut-off
Accumulated Evidence
Samples
(D)
Difference Cut-off
Accumulated Evidence
Samples
(E)
Accumulated Evidence
Samples
(F)
Accumulated Evidence
Samples
(G)
Exp. 1: r = 0.96, p < .001
RT(Number of frames)
Area under the DCB
(H)
Exp. 1: r = 0.8, p < .001
RT(Number of frames)
Area under the DCB

***Figure 2.*** *Model-free extraction of the decision boundaries. (A-B) Illustration of the boundary extraction for a decision simulated using a diffusion model with fixed (A) or collapsing boundary (B). The dashed black line corresponds to the original boundary with which the model was simulated and the red lines correspond to the model-free best-fitted DCB. The pink distributions correspond to data points within each trial in which the simulated participant continued sampling, whilst the blue and green distributions correspond to frames in which trials were terminated. The total area under the pink, blue and green distributions was normalized for each frame. (C-D) Same as (A-B), only for data that were simulated using the value cut-off (C) and difference cut-off heuristics (D). (E-F) Experimental data of a representative subject who participated in the numerical (E) and perceptual (F) experiments. Dashed black line corresponds to the model-based best-fitted boundaries, and the red lines correspond to the model-free best-fitted boundaries of the experiments. (G-H) Correlations between the area under the DCB and the mean RTs across participants in experiment 1 (G; $r = 0.96$, $p < 0.001$) and experiment 2 (H; $r = 0.8$, $p < 0.001$).*

The results are illustrated in Figure 2A-B, which shows that the DCBs (solid red lines) recover quite accurately the generating boundaries (dashed black lines). In particular, the extracted boundary is temporally constant or decreases as a function of time, when the generating boundary is flat or collapsing, respectively. Notably, the quantitative agreement between the generating and recovered boundaries is high. Note that the model-free boundary extraction method makes no a priori parametric assumptions about the shape of the model boundary. As shown in Figure 2C-D, for synthetic data generated using the non-integration-to-boundary (that is, value or difference cutoff heuristics), the three distributions of evidence (trials in which the model continues sampling (pink) and trials in which response has been made (blue and green)) show substantial overlap. Consequently, the decision classification curve algorithm fails to correctly classify the three classes of trials based on the integrated evidence. Thus, the presence of an accurate boundary classification curves (Figure 2A-B) provides strong support against (non-integration) cutoff models.

In both our datasets, we find stable classification curves (see Figure 2E-F, red lines, for representative example subjects) that imply a collapsing boundary, which are consistent with model fitting (Figure 2E-F, dotted black line), but make no parametric assumption on the form of this boundary (see Supplementary Figures 2 & 3 for the DCB of all participants). Finally, we linked the DCB to the basic behavioural measures of RT and accuracy. We found a high correlation between the area under the DCB and mean RT

across participants, in both experiment 1, $r = 0.96$, $p < 0.001$ (Figure 2G) and experiment 2, $r = 0.8$, $p < 0.001$ (Figure 2H). Correlations between the area under the DCB and accuracy were also found. However, they were weaker and less consistent (experiment 1: $r = 0.41$, $p = 0.03$, and experiment 2: $r = 0.29$, $p = 0.1$) (Supplementary Figure 4). Note that, although the correlation between RT and DCB is high, they are not interchangeable. An increased RT could also be caused by a reduction of the drift rate or by an increase in boundary without a time-collapsing shape.

So far, we have illustrated that the DCB can accurately reconstruct the decision boundary in the case of perfect evidence-integration. In the next section, we extend these results to the case in which the integration process is biased. We will show that, even then, the DCB can achieve three targets: (1) to indicate the presence of a bias, (2) to determine the simplest bias-free behaviourally equivalent model (see Supplementary Figure 5 for an illustration for the case of leaky integration, in which the DCB compensates for the evidence loss by a change only in the decision classification line) and (3) to obtain the actual boundary in the case of biased integration process, by selecting between a family of DCBs based on maximizing their classification performance metrics (see Supplementary Figure 6 for further details).

**Evidence integration is modulated by stimulus consistency.** A more detailed examination of the choice data in both experiments shows that stimulus consistency, operationalized as the absolute value of the difference between the number of frames with evidence favouring each alternative divided by the total number of frames in the trial (denoted as the difference in evidence directions (DED)), has a critical impact on choices and RT, above and beyond the effect of total evidence. To illustrate this, consider two trials with the same total evidence: trial 1 (2, 3, 1, 4, total evidence = 10) and trial 2 (6, -1, 8, -3, total evidence = 10). Whilst both trials have a total evidence of 10 in favour of one of the alternatives, the evidence stream is more consistent in trial 1 (consistency measure of $\frac{4-0}{4} = 1$) compared with trial 2, where half of the evidence favouring one alternative whereas the other half favors the other (consistency measure of $\frac{2-2}{4} = 0$).

**Table 1.** Beta coefficients for predicting accuracy, RT and confidence in experiment 1 and 2.

| | *β (S.E.)* | *t* | *p* | 95% *CI* |
|---|---|---|---|---|
| Experiment 1 | | | | |
| Accuracy | | | | |
| Evidence | 1.72 (0.06) | 28.03 | <0.001 | [1.60, 1.84] |
| Stimulus Consistency | 0.26 (0.05) | 5.69 | <0.001 | [0.17, 0.35] |
| RT | | | | |
| Evidence | -0.26 (0.01) | -28.23 | <0.001 | [-0.28, -0.24] |
| Stimulus Consistency | -0.14 (0.01) | -14.90 | <0.001 | [-0.16,-0.12] |
| Experiment 2 | | | | |
| Accuracy | | | | |
| Evidence | 0.56 (0.03) | 16.91 | <0.001 | [0.49, 0.62] |
| Stimulus Consistency | 0.27 (0.03) | 9.40 | <0.001 | [0.21, 0.33] |
| RT | | | | |
| Evidence | -0.16 (0.008) | -20.42 | <0.001 | [-0.18,-0.15] |
| Stimulus Consistency | -0.14 (0.008) | -16.78 | <0.001 | [-0.15,-0.12] |
| Confidence | | | | |
| Evidence | 0.13 (0.01) | 12.14 | <0.001 | [0.11, 0.15] |
| Stimulus Consistency | 0.15 (0.01) | 13.96 | <0.001 | [0.12, 0.17] |

*Results of mixed-model logistic regression for predicting accuracy (experiments 1 and 2) and mixed-model linear regressions for predicting RT (experiments 1 and 2) and confidence (experiment 2) using accumulated evidence and stimulus consistency as fixed factors and participants as random intercepts.*

Previous research has shown that stimulus consistency modulates decision confidence[41]. Here, we examine its impact also on accuracy and RT. To this end, we conducted several mixed-model regression analyses (logistic for accuracy and linear for RT and confidence), in which we predicted trial by trial choice accuracy, RT and confidence, using accumulated evidence and stimulus consistency (defined based on the evidence stream up to subject-initiated trial completion) as fixed factors and participants as random intercepts. The results

(Table 1) indicate that stimulus consistency improves accuracy and confidence and reduces RT, independent of the accumulated evidence.

Note that the DED is only one measure of stimulus consistency. More complex forms that include temporal factors can also be constructed. For example, the consistency bias may correspond to the size of the larger temporal cluster (LTC) with evidence in the same direction (that is, the largest cluster of evidence; see Supplementary Table 1 for analysis showing that such an LTC measure also predicts differences in accuracy, RT and confidence independent from total evidence).

**DCB modulation by stimulus consistency and model comparison.** Motivated by the results above, we examined whether stimulus-consistency (DED) modulates the DCB. To this end, we extracted the DCB of each frame whilst including the consistency measure (DED) as a predictor in the LDA model. Similar to the regression models above, we also included the mean evidence as a predictor in the model. We predicted that, if participants overweight consistent pieces of evidence, then less evidence will be required to reach a decision as consistency increases. Figure 3A-B show that this was indeed the case in experiments 1 and 2: as the consistency of the evidence increased, the DCB decreased (blue line), compensating for the bias in the evidence-integration process by setting a classification curve which is lower than the original boundary (orange line).

Note that this analysis does not provide a causal explanation for how consistency affects the evidence-integration process, but rather shows that it was biased by stimulus-consistency (see Discussion for details). This is because, the DCB as we showed so far, provides us with a behaviourally approximate (simpler, that is, without bias) model (Supplementary Figure 5). To provide a more mechanistic account, we will use two complementary methods and show that they converge to the same result. We will first rely on conventional model comparisons techniques to the select a model that provides the best fit for the data. Second, we will use a method developed to extract the decision-boundary using the DCB in case of biased integration process (Supplementary Figure 6). Using this method, we compare between a model assuming consistency-bias and a model which does not.

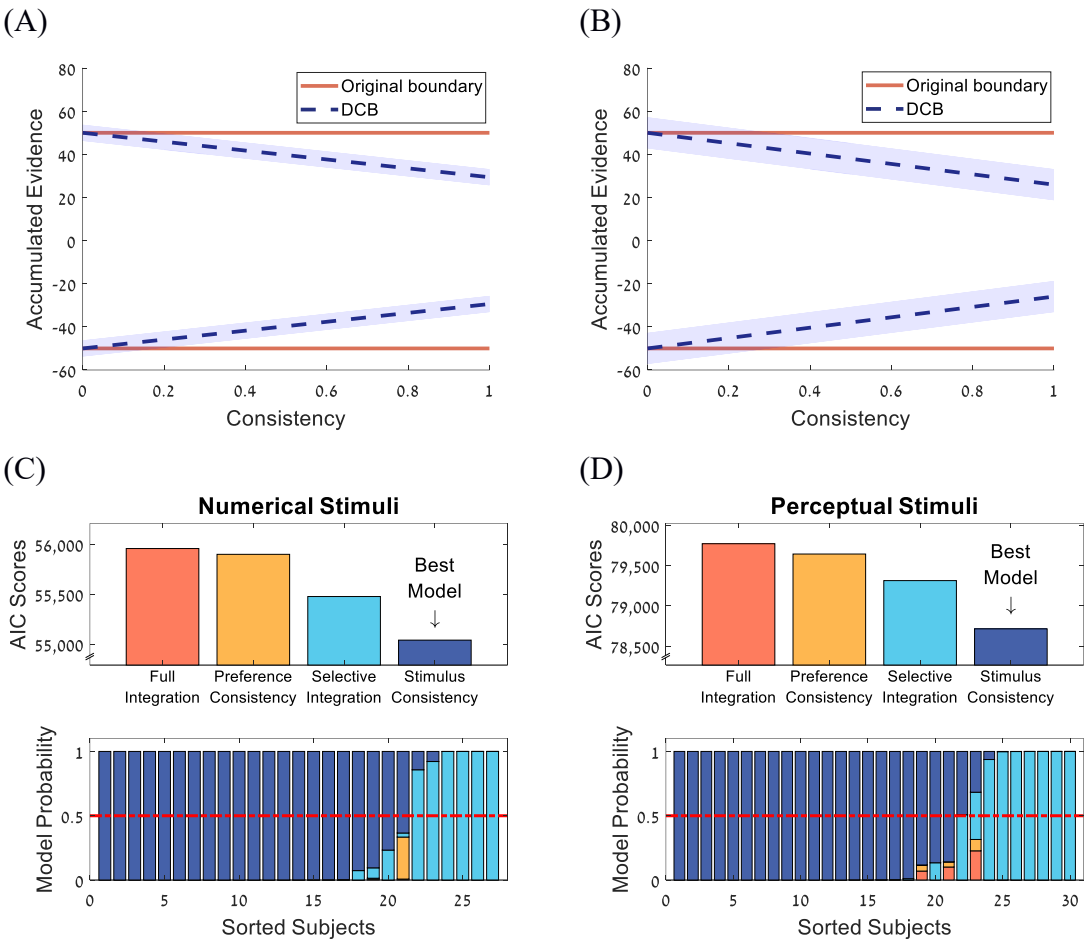



***Figure 3.*** *Results of experiments 1 and 2. (A-B) Modulation of the DCB as a function of consistency in experiment 1 with numerical stimuli (A) and experiment 2 with perceptual stimuli (B). In the case of a stimulus-consistency bias, the weight of consistent evidence is increased, thus less evidence is required to reach a decision. The DCB (dashed blue line) compensates for that by decreasing the classification curve with consistency, and having lower values compared to the original boundary (in orange). Note that, here, we used an arbitrary boundary value for illustration purposes and that the modulation of the DCB is averaged across participants. (C) Upper panel: The AIC group scores of the full-integration (orange), preference-consistency (yellow), SI (light blue) and stimulus-consistency (dark blue) models in experiment 1 with numerical stimuli. Lower panel: model probabilities for the individual participants. Colour coding is the same as in the upper panel. (D) As in (C) only for experiment 2 with perceptual stimuli.*

First, to specify the evidence-integration mechanism, we carried out a quantitative model comparison for each participant, using the Akaike Information Criterion (AIC) as a measure of fit, to include a model-complexity cost. We started with the perfect integration model (which assumes no systematic distortion of the evidence a subject integrates in each frame, that is, integration is only corrupted by additive noise). Next, we examined a selective-integration (SI) mechanism that amplifies or diminishes, respectively, the

stronger or weaker evidence within each frame across the two evidence streams[33,42] (see Computational methods). Critically, we also examined several variants of a stimulus-consistency model. In the simplest variant, the evidence is modulated solely based on whether evidence from the preceding frame is consistent (that is, it points in the same direction) with the current frame. In a slightly more complex variant, the modulation magnitude increases linearly with the number of consistent frames and resets to baseline (no modulation) with every swap (here, we report only the results of the more complex version, which provided better fit for the data). Finally, we also examined a preference-consistency model, wherein incoming evidence is modulated based on consistency with the total integrated evidence (thus reflecting momentary preference) up to that time point (see Computational methods for details).

In each model, we allowed for collapsing boundaries, which, as illustrated in Figure 2C-D, capture well the shape of the decision boundary (and provide much better fits to the data compared to fixed boundaries[34]). Figure 3C-D shows the group and individual participant fit measures for four models: full integration, selective integration, preference consistency and stimulus-consistency in experiments 1 and 2. As illustrated, the most successful of the models was the stimulus-consistency version in which evidence increased at each consecutive frame in the same direction, followed by the SI-model.

Finally, we developed a method to extracted the boundary using the DCB in the case of biased integration process (Supplementary Figure 6). This method uses a mixture of parametric and non-parametric methods. The former is used to characterize the biased integration process (but not the decision boundary), whilst the latter is used to extract the boundary using the DCB. To apply this method, a candidate biased integration model (for example, stimulus-consistency bias) is first selected. Then, instead of using the actual (unbiased) evidence to estimate the DCB, it is extracted based on the biased evidence (generated by simulating the biased integration model with different values of the bias parameter). For each level of the bias parameter, the DCB as well as a classification performance metric (for example, F1-score) are computed. This results in a family of DCB curves, one for each value of the bias parameter. The bias-parameter and corresponding DCB which maximizes accuracy (or any other performance metrics) comprise the estimate

(see Supplementary Figure 6 for further details and simulations). A bias-parameter that equals to 0 indicates that there was no bias in the evidence integration process (that is, full integration model), whereas a parameter higher than 0 indicates that the integration process was biased. Using this method, we found that the bias-parameter was higher than 0 for 81% of the participants in experiment 1 (mean bias parameter 2.26, 95% CI 1.70–2.86), and for 90% of the participants in experiment 2 (mean bias parameter 4.23, 95% CI 2.76–5.90), suggesting that in both experiment 1and 2, the integration process was biased by the consistency of the evidence.

**Experiments 3 and 4: testing the stimulus-consistency effects.** Because in our first two experiments participants' choices terminated the information stream, trial consistency depended, at least partially, on participants' responses and was not completely orthogonal to the integrated evidence. To address this limitation, and to directly test the impact of stimulus consistency on evidence-based choice and on decision confidence, two additional experiments were designed. In both, the number of samples was fixed and the stimulus consistency was manipulated completely independently from the total evidence. In experiment 3, sequences of eight number pairs were presented at a rate of 2Hz (as in experiment 1) and participants were instructed to choose at the end of the presentation the sequence with higher average (see Figure 4A). For each set of samples, we generated two paired trials, which consisted of the very same evidence content (across the eight time frames) for each sequence and differed only in temporal order of the values (that is, the paired trials varied in how the values on each side were shuffled). In the consistent trial condition (Figure 4C, bottom panel), one evidence stream provided stronger evidence in seven out of the eight frames, whereas in the inconsistent trial condition (Figure 4C, upper panel), each stream provided stronger evidence in four frames. In addition to consistency, the difficulty of the trials was also manipulated by sampling values from $X\sim N(52, 10^2)$ and $Y\sim(48, 10^2)$ for the difficult condition and from $X\sim N(52, 10^2)$ and $Y\sim N(44, 10^2)$ for the easy condition. In experiment 4, the stimulus-consistency effect was generalized to a much faster presentation rate of 12.5Hz and to a single stream of evidence (presented in the centre of the screen). Participants were presented with a sequence of eight arrays of red and blue dots (Figure 4B), and were instructed to determine whether more of the blue or the red dots were presented in total (see ref. [43] for similar paradigm). The number of dots presented in

each frame was sampled from a Normal distribution (corresponding to the differences between the left and right distributions in experiment 3): $X \sim N(2, 20^2)$ for the difficult trials and $X \sim N(5, 20^2)$ for the easy trials. As in Experiment 3, in the consistent condition, seven out of the eight frames provided support for one of the alternatives and in the inconsistent condition each alternative was supported by four of the frames. Critically, consistency and difficulty were manipulated completely orthogonally.

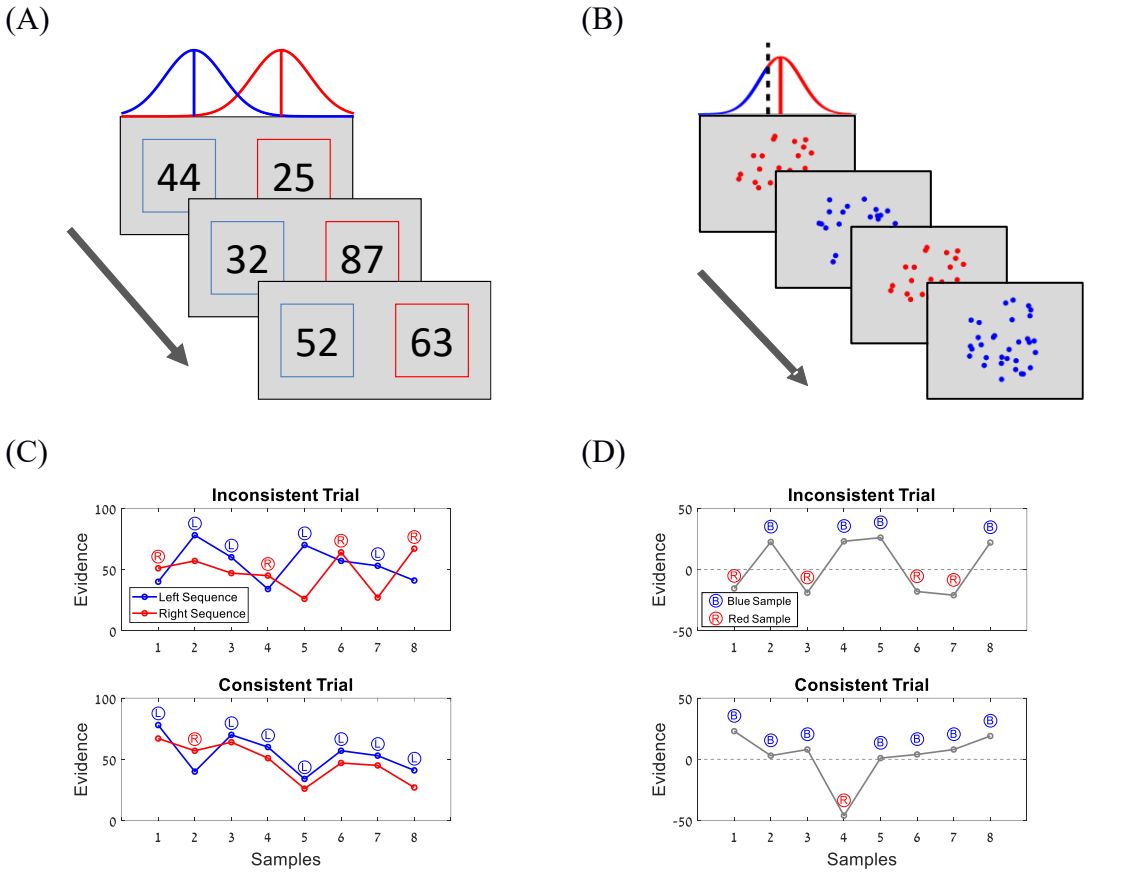



***Figure 4.*** *Experimental paradigms in experiments 3 and 4. (A) Participants in experiment 3 were presented with pairs of numerical values sampled from two overlapping Normal distributions (as in experiment 1), and were asked to choose which sequence was drawn from a distribution with a higher mean. The presentation was terminated after eight pair values (that is, an interrogation protocol). (B) Participants in experiment 4 were presented with a single stream of eight arrays of blue and red dots, and were asked to indicate whether more blue or red dots were presented in total. The dashed black line indicates a value of 0. (C) Illustration of consistent and inconsistent trials in experiment 3. Note that both trials have exactly the same amount of evidence. 'L' and 'R' symbols correspond to momentary advantages of the left or right sequences, respectively. (D) Illustration of consistent and inconsistent trials in experiment 4. 'R' and 'B' symbols correspond to red and blue samples, respectively. As in experiment 3, both trials have exactly the same amount of evidence.*

As shown in Figure 5A-D, the participants were both more accurate and more confident in consistent trials than in inconsistent trials, for both the easy and difficult conditions, in experiment 3 (accuracy/easy: permutation test $p < 0.001$, Cohen's $d = 1.26$, 95% CI 0.06 – 0.13, accuracy/difficult, permutation test $p < 0.001$, Cohen's $d = 1.40$, 95% CI 0.11 – 0.20, confidence/easy: permutation test $p < 0.001$, Cohen's $d = 1.11$, 95% CI 0.07 – 0.14, confidence/difficult: permutation test $p < 0.001$, Cohen's $d = 0.94$, 95% CI 0.04 – 0.10) and experiment 4 (accuracy/easy: permutation test $p < 0.001$, Cohen's $d = 1.69$, 95% CI 0.12 – 0.18, accuracy/difficult: permutation test $p < 0.001$, Cohen's $d = 2.19$, 95% CI 0.22 – 0.31, confidence/easy: permutation test $p < 0.001$, Cohen's $d = 1.56$, 95% CI 0.11 – 0.19, confidence/difficult: permutation test $p < 0.001$, Cohen's $d = 1.49$, 95% CI 0.11 – 0.19). Interestingly, in both experiments, the modulation of the confidence responses was different for correct and incorrect responses. Whereas for correct responses, confidence increases with stimulus consistency in experiment 3 (correct/easy: permutation test $p < 0.001$, Cohen's $d = 0.94$, 95% CI 0.05 – 0.13, correct/difficult, permutation test $p < 0.001$, Cohen's $d = 0.80$, 95% CI 0.03 – 0.10) and experiment 4 (correct/easy: permutation test $p < 0.001$, Cohen's $d = 1.58$, 95% CI 0.11 – 0.18, correct/difficult, permutation test $p < 0.001$, Cohen's $d = 1.41$, 95% CI 0.11 – 0. 18), this pattern was not obtained for incorrect trials in experiment 3 (incorrect/easy: permutation test $p = 0.53$, Cohen's $d = -0.14$, 95% CI -0.05 – 0.02, incorrect/difficult: permutation test $p = 0.96$, Cohen's $d = 0.01$, 95% CI -0.03 – 0.03) or experiment 4 (incorrect/easy: permutation test $p = 0.19$, Cohen's $d = -0.29$, 95% CI -0.14 – 0.01, incorrect/difficult: permutation test $p = 0.37$, Cohen's $d = -0.20$, 95% CI -0.11 – 0.03). These findings indicate that meta-cognitive accuracy (confidence-resolution = $\text{confidence}_{\text{corrects}} - \text{confidence}_{\text{errors}}$) increases with stimulus consistency in experiment 3 (resolution/easy: permutation test $p < 0.001$, Cohen's $d = 0.82$, 95% CI 0.05 to 0.16, resolution/difficult: permutation test $p = 0.01$, Cohen's $d = 0.59$, 95% CI 0.02 to 0.11) and experiment 4 (resolution/easy: permutation test $p < 0.001$, Cohen's $d = 0.82$, 95% CI 0.10 to 0.29, resolution /difficult: permutation test $p = 0.001$, Cohen's $d = 0.73$, 95% CI 0.09 to 0.27). As shown in Figure 5C-D (see also Supplementary Figure 11), this effect can also be accounted for by the stimulus-consistency model by applying a signal-detection confidence approach.

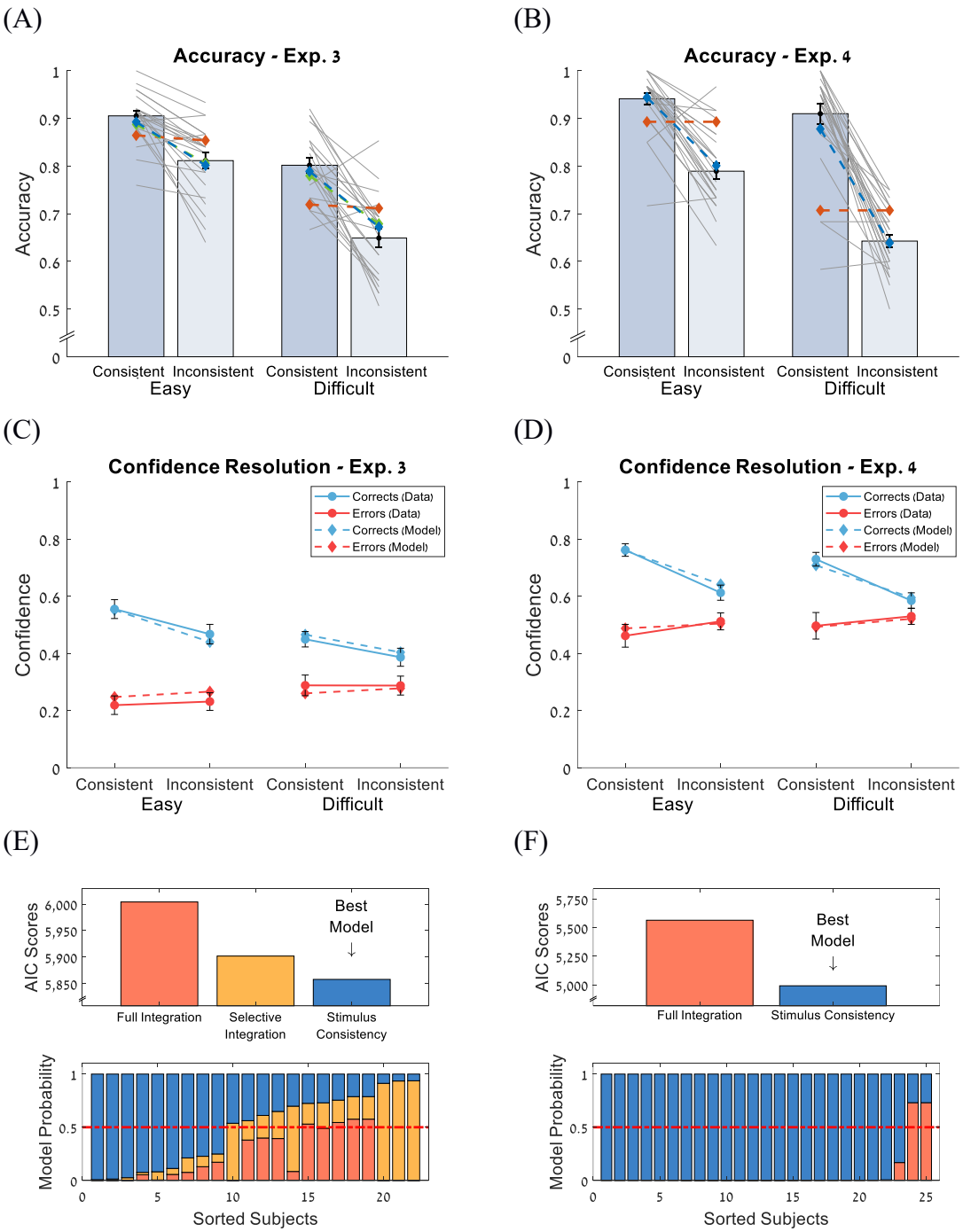



***Figure 5.*** *Results of experiments 3 and 4. (A) Choice accuracy in experiment 3 as a function of difficulty (separation between the sampling distributions: easy versus difficult) and consistency (difference in the directions of the evidence: consistent versus inconsistent). The red, green and blue lines correspond to the predictions of the full/leaky-integration, SI and stimulus consistency models, respectively. The thin grey lines correspond to individual participants (n = 22 participants). (B) Same as (A) for experiment 4. Note that, here, the predictions of the SI model were not included as only one stream of evidence was presented (n = 25 participants). (C) Confidence as a function of difficulty and consistency for correct (blue lines) and incorrect (red lines) responses. Data are*

*shown with solid lines and circle symbols. Model predictions are shown with dashed lines and diamond symbols (n = 22 participants). (D) Same as (C) for experiment 4 (n = 25 participants). (E) Model comparison for experiment 3. The stimulus consistency model outperformed the leaky and SI models. (F) Model comparison results for experiment 4. As in experiment 3, the stimulus consistency model outperformed the integration model. Data are presented as mean values ± standard error of the mean.*

We conducted quantitative model comparison for the several types of integration models using the data from experiments 3 and 4. Since the data of experiment 3 showed a recency pattern, we used leaky[10,44,45] instead of full integration as our default model for that experiment (Computational methods). Overall, we compared the following models: (1) leaky/full-integration models, in which there is no distortion of the integrated evidence other than the decaying temporal weights, (2) SI model, which, additional to integration leak, gives higher weight to high values compared to low values, within each frame (note that as only one stream of evidence was presented in experiment 4, the SI model (which assumes weighting based on the comparison between pairs of samples) was excluded from the model comparison of that experiment) and (3) stimulus-consistency model (from experiment 1-2; Figure 3C-D). The results show that the stimulus-consistency model outperformed the other models, in both experiment 3 and in 4, and provides the best account for the data at group levels as well as for the majority of participants (Figure 5E-F).

Interestingly, whereas both of the bias models are able to account for the modulation of accuracy by stimulus consistency in experiment 3 (Figure 4C), the stimulus-consistency model accounts for subtler temporal clustering effects in the data. For example, unlike the SI model, the stimulus-consistency model predicts that accuracy is modulated by the size of the largest cluster of evidence consistent with the correct choice (LTC; see Supplementary Table 1 for the impact of this measure in experiments 1 and 2 and Supplementary Figure 10 for data showing an association between the LTC-enhancement and the advantage of stimulus-consistency over the SI model in experiment 3).

**Stimulus-consistency and normativity.** Why do participants increase the relative weighting of pieces of evidence that are consistent with previous ones? At face value, this distortion introduces a gap between the accumulated and 'real' evidence and should reduce task performance. Indeed, this is the case in the absence of (or for low) integration noise. As illustrated in Fig, 6A-B, however, in the presence of high integration noise, the consistency modulation makes the mechanism more robust to the corrupting impact of this noise (see cross-over between red and light blue lines in Figure 6A, so that, for each level of noise, there is a consistency-modulation that optimizes performance; black dots in Figure 6B and dark blue line in Figure 6A). A similar robustness effect due to selective integration was reported for the SI-model[33,42] (see also ref. [42]). In both cases, the integration mechanism distorts the actual evidence by shifting the distribution of accumulated evidence towards the correct side (see centre of the blue/red Gaussians in Figure 6C-D, whilst also making this distribution broader). Whilst for low noise this is detrimental to performance, for high noise it is beneficial, as the shift helps to make the effect of additional noise less pronounced. Note that, since we are looking at normative considerations, the current simulations exclude integration leak. Adding it to all models does not change any of the results.

## Discussion

In the present study, we examined the mechanism that human observers deploy when making decisions on rapid streams of (perceptual and numerical) stochastic evidence. Using a behavioral model-agnostic method (the DCB curve), we showed that decision making is well characterized by integration to boundary rather than by non-integration heuristics, and that the boundary collapses with the passage of time (Figure 2E-F; see also ref. [46]). Furthermore, we found that DCB curves constitute an informative behavioral benchmark for evaluating biases in evidence-integration. In particular, they provide a simplest bias free evidence integration model that approximates a biased integration model (Supplementary Figure 5). The covariation of the DCB with stimulus-consistency (Figure 3A-B) indicates a consistency bias in the evidence integration.

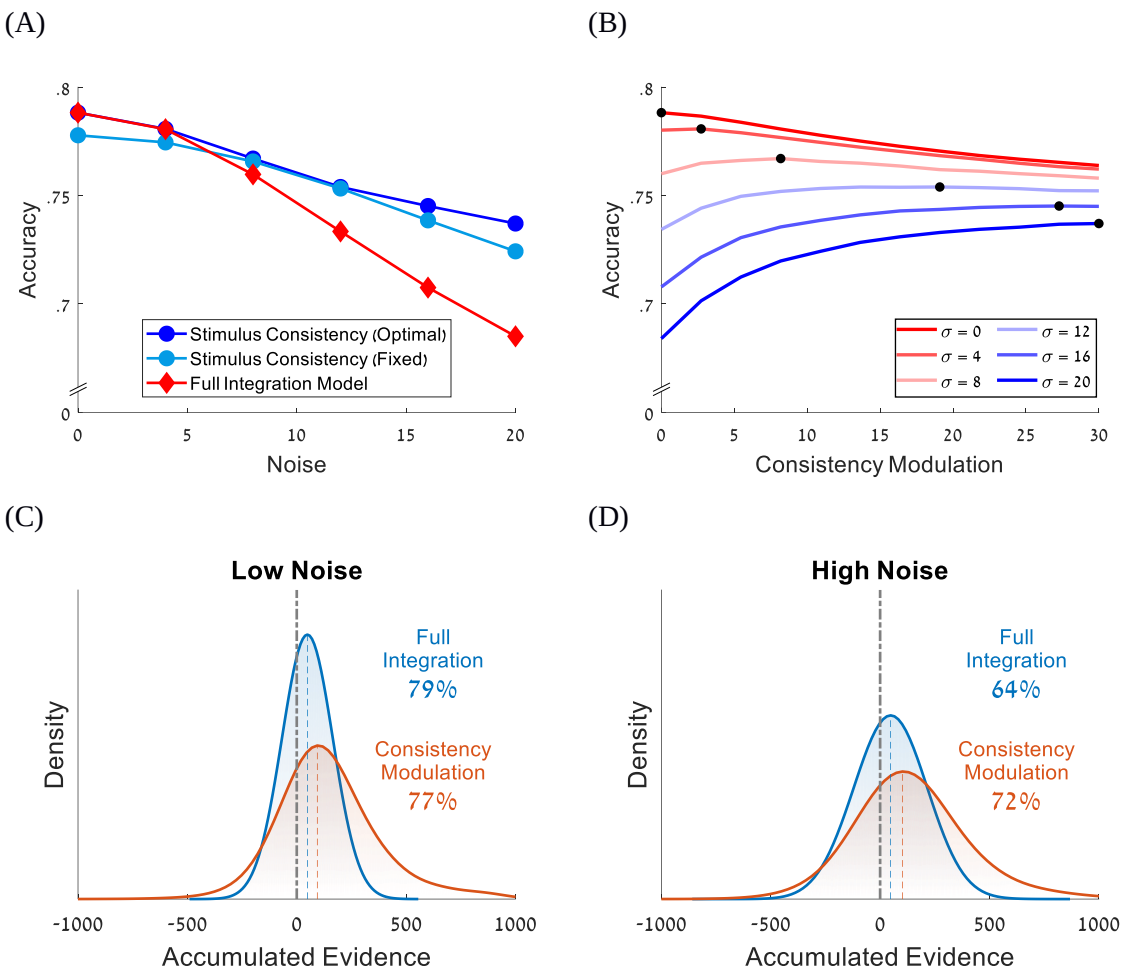



***Figure 6.*** *Stimulus-consistency and normativity. (A) accuracy (that is, robustness to noise) as a function of noise, separately for the full integration (red line) and stimulus consistency models (blue lines). The accuracy of the stimulus-consistency model is presented for the model simulated using a fixed consistency parameter value of 10 (light blue) and for the model simulated using the optimal consistency parameter for each noise level (dark blue; see also B). (B) Accuracy of the stimulus-consistency model as a function of the consistency-modulation strength, for different levels of noise (σ curves). Black circles indicate consistency-values that maximize accuracy for a given level of accumulation noise. One can see that, as the level of the noise increases (red to blue lines), so does the level of the consistency modulation needed to achieve optimal performance (black circles). (C-D) The distributions of the total accumulated evidence at the moment of response for the full integration model (blue) and stimulus-consistency model (red). The accuracy of the full integration model is higher than that of the stimulus-consistency model for low noise simulations (C), but the stimulus-consistency model is more accurate for high-noise simulations (D).*

The DCB receives as input the integrated evidence excluding internal noise (Eq. 2). As a result, its success in recovering the decision boundary is limited to experimental designs in which stimulus variability is high enough compared to internal noise. Consequently, the DCB method becomes less reliable when the presentation rate becomes closer to the integration time constant (which we assume to correspond to about 30ms[47,48]). Supplementary Figure 7 shows the results of a simulation examining the accuracy of the DCB method as a function of presentation time of each frame. As shown, the boundary reconstruction error increases as the presentation time of each frame decreases. This occurs because, when the evidence samples are presented close to the visual integration time scale, the neural responses to consequent sample become fused and the therefore stimulus variability is decreased. Whilst this could make the method difficult to apply to some prominent tasks such as the randomly moving dots[49,50], we believe than a rate of 5-12.5 Hz (as used here) is a reasonable one to most ecological tasks in which subjects make decisions based on stochastic sequences of evidence[43,51] and which sets the evidence integration at the cognitive rather than perceptual level.

Motivated by previous studies, we examined here two types of evidence-integration biases. The first is an attention bias, which affects the relative weight of evidence given to temporally simultaneous sources of evidence[26]. In the SI model, for example, the higher of the two values presented on each frame, receives a higher weight than the lower one[23,52,53]. The second bias involves the sequential impact of a frame on subsequent frames, whereby evidence that is consistent with previous frames receives higher weight than inconsistent evidence. Model comparisons, supported the consistency bias in accounting for our data. Notably, consistency affected not only decision accuracy, but also decision confidence[41], such that consistent evidence facilitated high decision confidence even after controlling for the total amount of evidence. Critically, consistency did not merely exert a biasing influence on confidence but improved participant's meta-cognitive performance as measured by the resolution of confidence (that is, the correlation between confidence and choice correctness). Indeed, confidence as a function of consistency increased for correct choices but remained constant for incorrect choices. An open question for future studies is whether the effects that consistency exerts on choice accuracy and meta-cognition are dissociable.

Previous research has reported sequential effects operating at the trial level. For example, a choice biases the interpretation of evidence in subsequent trials[54–56]. Similarly, a preliminary decision biases processing of additional post-choice evidence towards confirming the initial decision[28,31,57], and decisions bias the strength-evaluation of pro-choice evidence that led to it[58–60]. Common to these studies is the assumption that these biases are triggered by the formation of a decision. Our findings, however, extend this notion by suggesting that a similar micro-level evidence integration bias operates during decision formation, before committing to a choice (see also refs. [61,62]). In particular, we found that the best-fitting model was one in which the evidence is boosted for consistent evidence frames (this boost increases with the number of consecutive consistent frames) and is reset to baseline when the first inconsistent piece of evidence is encountered. Despite its evidence distortion, we have shown that this mechanism has an adaptive function in the presence of integration noise. Since evidence samples that are consistent with their predecessors are more likely to carry stronger evidence in favour of the correct alternative, inflating their weight provides extra protection from the corrupting effect of accumulation noise, resulting in increased decision accuracy (Figure 6). This 'normativity hypothesis' predicts that, to the extent that consistency-based evidence integration is a controlled and adjustable strategy, consistency sensitivity effects will increase as a function of integration noise, for example, when one is performing a dual task or when working memory is loaded. We leave this interesting question for future studies (see ref. [32] for a parallel investigation in relating to selective-integration).

Our findings raise the intriguing hypothesis that confirmation biases are a form of consistency bias, whereby post-choice evidence inconsistent with pre-choice evidence is integrated less effectively than post-choice consistent evidence. Future studies should investigate whether, to what extent, and how consistency and confirmation biases are related, by measuring both biases within participants and using a unified paradigm. Another interesting possibility is that the consistency of evidence supporting a decision might affect the extent of a confirmation bias. For example, choices that are based on more consistent evidence may likely be more prone to confirmation bias, for example, due to the mediating effect of decision confidence[30]. Future studies could also beneficially examine whether and how consistency bias is related to a broad range of individual traits such as the need for

cognitive closure[63], political radicalism[57] and dogmatism[64], or to psychiatric conditions such as OCD[65]. Consistency bias may be also related to higher-level judgments such as legal decisions, in which overweighting of consistent evidence may lead to erroneous decisions. We believe our paradigm provides an important advantage over current confirmation bias paradigms by addressing these questions. In current confirmation bias paradigms, which probe one's ability to revise initially wrong decisions, validity might be jeopardized by demand characteristics (for example, presenting oneself in a self-consistent manner). In contrast, the current approach eschews these concerns, since participants are not required to contradict or confirm their previous decisions.

In conclusion, our methods allowed us to validate critical aspects of the evidence accumulation process and to unravel the biases that affect it. Our findings contribute to a growing literature speaking to the notion that self-inflicted distortions of evidence are ironically adaptive in that they act to increase choice veracity by making it robust to noise. A critical next step is to study how these strategies are acquired and how they relate to puzzling behaviors such as confirmation bias, and to better characterize the environmental and psychological variables that affect strategy selection.

## Methods

**Experimental methods.** *Participants.* The participants in experiments 1, 2 and 3 were undergraduates from Tel Aviv University: $n$ = 27 (22 female, age 21–28 years) in experiment 1 (data taken from ref. [34]), $n$ = 30 (22 female, age 18–35 years) in experiment 2 and $n$ = 22 (17 female, age 21–30 years) in experiment 3. The participants in experiment 4 were recruited via Prolific (https://prolific.ac/), $n$ = 25 (12 female, age 18–32 years). All participants reported having normal or corrected-to-normal vision. The participants in experiments 1, 2 and 3 received course credit in exchange for taking part in the experiments, as well as a bonus fee ranging from 15 to 25 ILS, which was determined by their task performance. The participants in experiment 4 received £4 in exchange for participation. All experiments were approved by the ethics committee of Tel Aviv University.

*Stimuli.* The stimuli consisted of pairs of numerical values (experiments 1 and 3) or bars (experiment 2) which were presented simultaneously (Figure 1A-B), at a rate of 2 Hz (500ms per frame, experiment 1 and 3) or 5 Hz (200ms per frame, experiment 2). In experiment 4, the stimuli consisted of a single stream of red and blue arrays of dots (Figure 4B) presented at a rate of 12.5 Hz (80ms per frame). Displays in experiments 1 and 3 were generated by an Intel I7 personal computer attached to an Asus 24" 248qe monitor with a refresh rate of 144 Hz, using the 1920×1080 resolution graphics mode. Displays in experiment 2 were generated by an Intel I3 personal computer attached to a ViewSonic 19" Graphics Series G90fB CRT monitor with a refresh rate of 60 Hz using the 1024×768 resolution graphics mode. Experiment 4 was designed in PsychoPy3[66] and hosted online using Pavlovia (http://www.pavlovia.org/). Responses were collected via the computer keyboard. The viewing distance was approximately 60 cm from the monitor.

*Task and Design.* Each trial in the experiments began with a fixation display consisting of a black 0.2° × 0.2° fixation cross (+) that remained on the screen for 1s. Then, pairs of numerical values (experiments 1 and 3), bars (experiment 2) or arrays of dots (experiment 4) were presented sequentially to the participants, who were asked to decide which sequence was drawn from a distribution with a higher mean value (experiment 1), greater mean length (experiment 2), which of the sequences has a higher mean (experiment 3) or whether more blue or red dots were presented in total (experiment 4). The presentation in experiments 1 & 2 was terminated by the participants' response (free response protocol), whilst the presentation in experiments 3 and 4 was terminated after eight samples (interrogation protocol). In experiment 1, trials in which the response was faster than 250ms or was performed after more than 11 samples were presented were excluded from further analysis (less than 2% of the data). In experiment 2, trials in which the response was faster than 200ms, performed after more than 20 samples were presented, or in which choice or decision confidence were not recorded were excluded from further analysis (less than 5% of the data). Responses were given by pressing the arrow keys (experiments 1 and 2; left/right arrow keys for the left/right sequences, respectively) or by using the computer mouse (experiments 3 and 4). In experiments 2, 3 and 4, after each choice, participants were also asked to indicate their choice confidence. In experiment 2, we used a continuous scale with end points labeled '50%' and '100%'. In experiment 3, we used a six-button radio

scale with end points (that is, 1 and 6) labelled 'Not confident at all' and 'Very confident'. In experiment 4, we used a continuous scale with end points labelled 'Not confident at all' and 'Very confident'. Confidence scores were normalized using min-max normalization:

$$Normalized\ Confidence = \frac{confidence\ - \min(confidence)}{\max(confidence)\ - \min(confidence)}$$

*Experimental Conditions.* In all the experiments, the samples were drawn from Gaussians distributions. Experiment 1 included two difficulty levels, which were manipulated by varying the separation between the Gaussians: in the easy trials the means of the Gaussians were: $\mu_1$=52 vs. $\mu_2$=44, $\sigma$=10, whilst in the difficult trials the means were: $\mu_1$=52 vs. $\mu_2$=48, $\sigma$=10. Experiment 2 also consisted of two difficulty levels with $\mu_1$ = 52.5 vs. $\mu_2$ = 47.5 and $\mu_1$ = 51.5 vs. $\mu_2$ = 48.5, as well as an orthogonal manipulation of the sequences variance with $\sigma_1$ = 0.1167 vs. $\sigma_2$ = 0.07. In experiment 3, we orthogonally manipulated the difficulty and the consistency of the evidence. Difficulty was manipulated by increasing the separation between the Gaussians from $\mu_1$=52 vs. $\mu_2$=48, $\sigma$=10 (difficult trials) to $\mu_1$=52 vs. $\mu_2$=44, $\sigma$=10 (easy trials). Consistency was manipulated by sampling eight values from the high, as well as from the low mean distributions. Then, to generate consistent trials, we paired these values such that, in seven out of the eight pairs, the stronger evidence was in favour of the higher-mean distribution (Figure 4C, lower panel). To generate inconsistent trials, we shuffled the temporal order of the same values and repaired them such that in only four out of the eight pairs was the stronger evidence in favor of the higher mean distribution (Figure 4C, upper panel). In experiment 4, difficulty and the consistency were also orthogonally manipulated. Difficulty was manipulated by decreasing the mean of the Gaussian from $\mu$=5 to $\mu$=2 ($\sigma$=20 in both conditions). Consistency was manipulated by generating two types of trials: consistent trials, in which seven out of the eight frames provided support for one of the alternatives (Figure 4D, lower panel), and inconsistent trials, in which each alternative was supported by four of the frames (Figure 4D, upper panel).

**Statistical analysis.** *Correlations and mean comparisons.* Correlations were examined using Pearson correlation coefficient. Means were compared using permutation tests with$10^5$ random shuffles. All tests were two-sided.

*Mixed effects models.* The effect of accumulated evidence and stimulus consistency on accuracy, RT confidence in experiments 1 and 2 (Table 1) were estimated on a trial-by-trial basis using mixed model regression analyses. The regressions were implemented using MATLAB 'fitlme' and 'fitglme' functions with participants serving as random effects and with a free covariance matrix. The fixed effects variables were: (1) accumulated evidence, calculated as the sum of differences between the two streams of evidence at the moment of response, and (2) stimulus-consistency, calculates as the absolute value of the difference between the number of frames with evidence favouring the two alternatives, normalized by the length of trial. Both variables were normalized using z-score transformations.

Choices in Experiment 1 & 2 (coded as 1 for correct and 0 for error) were predicted using logistic regressions, which in Wilkinson notation was:

$$logit(P_{Choices}) \sim (accumulated\ evidence) + (stimulus\ consistency) + (1|subject)$$

RTs (experiments 1 and 2) and confidence (experiment 2) were predicted using linear regressions, which in Wilkinson notation were:

$$RT \sim (accumulated\ evidence) + (stimulus\ consistency) + (1|subject)$$
$$Confidence \sim (accumulated\ evidence) + (stimulus\ consistency) + (1|subject)$$

The exact same pattern of results reported in Table 1 was obtained if the accumulated evidence and stimulus-consistency were also included as random effects.

**Computational methods.** *Model-free method*. The validity of the model-free method was tested by simulating 10,000 synthetic decisions using known (fixed or collapsing) boundaries and examining the ability of the model-free method to accurately recover them. The values in all simulations (Figure 2A-D) were sampled from $X \sim N(52, 15^2)$ and $Y \sim N(46, 15^2)$. The decision process in Figure 2A-B was based on Eq. 1 with either a fixed (Figure 2A) or collapsing boundary (Figure 2B). The fixed-boundary was characterized by a single boundary parameter ($c$=50), and the collapsing-boundary was characterized by four parameters: intercept, shape, scale and asymptote ($a$=100, $k$=3, $\lambda$=4, $a'$=10; see Modelling

details and ref. [38]). The decision process in Figure 2C-D was based on the value and difference cut-off heuristics with cut-offs of 70 and 20, respectively[34].

The decision boundaries in Figure 2A-D were extracted by applying the LDA algorithm[39,40] to the integrated evidence excluding internal noise (that is, *Y(t)*, see Eq. 2). For each frame (*t*), each trial that was not terminated before *t*, was classified as one of the following categories: choose alternative A, choose alternative B or continue sampling. Then, using the LDA, we extracted the planes that optimize the separation between different classes for each frame. We assumed that the upper and lower boundaries are symmetrical and therefore averaged both.

As mentioned in the main text, the internal noise causes an unavoidable overlap between the different classes. This overlap impairs the LDA ability to correctly extract the decision boundary and is particularly evident in slow trials due to the accumulation of internal noise across time[67]. Thus, to increase the robustness of the model-free method to internal noise, we constrained the boundary extraction of each frame by previous ones. To this end, we extracted the boundary based on two predictors: $t$ and $Y(t)$, $1 \leq t \leq n$, as illustrated in the table below:

**Table 2 Illustration of the data used by the LDA algorithm**

| Y(t) | t | Response classification |
|---|---|---|
| $Y_{\text{trial 1}}$ (t = n) | n | Choose B |
| $Y_{\text{trial 3}}$(t = n) | n | Continue sampling |
| … | … | … |
| $Y_{\text{trial 1}}$ (t = n-1) | n-1 | Continue sampling |
| $Y_{\text{trial 2}}$ (t = n-1) | n-1 | Choose A |
| $Y_{\text{trial 3}}$ (t = n-1) | n-1 | Continue sampling |
| … | … | … |
| $Y_{\text{trial 1}}$ (t = 1) | 1 | Continue sampling |
| $Y_{\text{trial 2}}$ (t = 1) | 1 | Continue sampling |
| $Y_{\text{trial 3}}$ (t = 1) | 1 | Continue sampling |
| … | … | … |

*The LDA classifies each frame to one of three classes: terminating the trial by choosing alternative A, terminating the trial by choosing alternative B or continue sampling more evidence, based on the time point (t) and the evidence accumulated to this time point (Y(t)).*

The LDA algorithm provide linear functions that separate the different classes from each other. To obtain the value of the boundary for the n-th frame, we computed the value of the separating linear functions for this frame.

**Modelling Details.** *Integration to boundary models.* We examined several integration-to-boundary models, all of them assumed integration of evidence based on the following formula:

$$X(t) = X(t-1) + \mu(t) + \varepsilon(t),\ \varepsilon \sim N(0, \sigma^2) \qquad \text{Eq. (1)}$$

where *X(t)* is the accumulated differences at time *t* and $\varepsilon(t)$ is a random Gaussian noise, which is independent from the evidence-sampling noise. The term $\mu(t)$ varied between the different models as follows:

1) Full integration

$$\mu(t) = V_L(t) - V_R(t) \qquad \text{Eq. (3)}$$

where $V_L(t)$ and $V_R(t)$ are the samples drawn from the left and right distributions at time *t*, respectively (note that $V_L(t)$ and $V_R(t)$ include the sampling noise).

2) Stimulus consistency

$$\mu(t) = V_L(t) - V_R(t) + sign(V_L(t) - V_R(t)) \cdot \theta \cdot i \qquad \text{Eq. (4)}$$

where $\theta$ is a free parameter representing the enhancement given to pieces of evidence that are consistent with previous ones and $i$ counts the run of consistent values (starting at 0). For example, if the differences between the values are: 15, 20, 8, -15 and -25, then $i_{t=1} = 0, i_{t=2} = 1, i_{t=3} = 2, i_{t=4} = 0$ and $i_{t=5} = 1$.

3) Preference consistency

$$\mu(t) = \begin{cases} (V_L(t) - V_R(t)) \cdot \theta, if\ \dfrac{sign\left(V_L(t) - V_R(t)\right)}{sign\left(X(t-1)\right)} = 1 \\ V_L(t) - V_R(t), otherwise \end{cases} \qquad \text{Eq. (5)}$$

where $\theta$ is a free parameter representing the enhancement given to pieces of evidence consistent with the total accumulated evidence at time $t-1$.

4) Selective integration

$$\mu(t) = V_L(t) \cdot \frac{1}{1+e^{-\theta(V_L(t)-V_R(t))}} - V_R(t) \cdot \frac{1}{1+e^{-\theta(V_R(t)-V_L(t))}} \quad \text{Eq. (6)}$$

where $\theta$ is a free parameter affecting the magnitude of the selective gating[23].

All the integration models in experiments 1 and 2 assume integration to a collapsing-boundary[34], modelled using a Weibull cumulative distribution function[38]:

$$u(t) = a - [1 - exp\left(-\left(\frac{t}{\lambda}\right)^k\right)] \cdot (a - a') \quad \text{Eq. (7)}$$

where $\pm u(t)$ are the upper/lower thresholds at time $t$, $a/a'$ are the initial (intercept) and asymptotic values of the boundary, respectively, and $\lambda$ and $k$ are the scale and shape parameters of the Weibull function, respectively.

In experiments 3 and 4, we used an interrogation paradigm, in which the probability of choosing each alternative was calculated using an exponential version of Luce's choice rule[68]:

$$P(Left) = \frac{1}{1+e^{-\left(\beta_0+\beta_1\left(\sum_{t=1}^{n}\mu(t)\right)\right)}} \quad \text{Eq. (8)}$$
$$P(right) = 1 - P(Left)$$

where $\beta_1$ indicates the sensitivity of the model to the accumulated evidence, with an intercept of $\beta_0$.

In addition, we examined whether the participants in experiments 3 and 4 showed a recency bias, as reported in several previous studies which used an interrogation paradigm[33,53]. To this end, for each participant, we performed a temporal logistic regression analysis, in which we predicted the response of each trial based on the differences between $V_L$ and $V_R$ at each frame, ranked by their temporal order. In 3xperiment 3, the mean weight of samples 5-8 was significantly higher than that of samples 1-4 (permutation test $p < 0.001$, Cohen's

d = 2.09, 95% CI 0.30 – 0.45) indicating a recency bias. No such effect was found in experiment 4 ($p$ = 0.09, Cohen's d = 0.35, 95% CI -0.01 – 0.10). This motivated us to include a leak term in experiment 3[10], which controls the extent to which earlier values are given less weight. Thus, Eq. 8 was extended to the following form in Experiment 3:

$$P(Left) = \frac{1}{1 + e^{-\left(\beta_0 + \beta_1\left((1-\lambda)^{n-t} \cdot \sum_{t=1}^{n} \mu(t)\right)\right)}} \qquad \text{Eq. (9)}$$
$$P(right) = 1 - P(Left)$$

where $\lambda$ is the leak term.

*Non-integration to boundary models.* We examine three models that did not assume integration of evidence decision boundary (Figure 1C-D and Figure 2C-D). The first model is the value-cutoff heuristic, which assumes that observers choose based on the detection of a single high-value sample. For example, if a participant uses a cut-off value of 70, then they will choose the sequences in which a value higher than 70 first appears. The second heuristic is the difference-cutoff heuristic, which assumes that observers choose based on the first frame in which the difference between the numbers exceeds a predetermined threshold[34]. In addition to these two heuristics, we examined a third model which we labeled a random-timer model. This model assumes integration of evidence based on Eq. 1, but the RT is determined by an exogenous process. The value cut-off heuristic was simulated using thresholds of 70 (experiment 1, Figure 1C), 80 (experiment 2, Figure 1D) and 70 (DCB analysis, Figure 2C). The difference cut-off heuristic was simulated using thresholds of 20 (experiment 1, Figure 1C), 25 (experiment 2, Figure 1D) and 20 (DCB analysis, Figure 2D). The RTs of the random timer model were sampled from an ex-Gaussian distribution with $\mu = 3$, $\sigma = 0.5$ and $\lambda = 2/3$ (experiment 1, Figure 1C) and $\mu = 6$, $\sigma = 1$ and $\lambda = 3$ (experiment 2, Figure 2C). These values were chosen because they provided accuracy and RTs similar to the ones observed in the data of experiments 1 and 2.

*Optimization procedure.* The free parameters of the computational models were fitted to the data (choices and decision times) of each participant in experiments 1 and 2 separately, using maximum likelihood estimation. For each trial, we simulated the different models 1,000 times for a given set of proposal parameters and calculated the proportion of trials in which the model choice and decision time matched the empirical data. Denoting the

proportion of match between the simulated and empirical data by $p_i$, we maximized the likelihood function $L(D|\theta)$ of the data ($D$) given a set of proposal parameters ($\theta$), by:

$$L(D|\theta) = \prod_{i=1}^{N} p_i$$

To find the best set of proposal parameters we first used an adaptive grid search algorithm (see ref. [69] for details) and then used the three best sets of proposal parameters as starting points to a Simplex minimization routine[70]. This data-fitting procedure showed good to excellent ability[71] to recover the free parameters of the models (see Supplementary Figures 8 & 9).

**Data availability**

The data that support the findings of this paper are available at https://osf.io/vywbx/

**Code availability**

The codes used for all studies are available at https://osf.io/vywbx/

**Acknowledgments**

This research was supported by a grant to Marius Usher from the Israel Science Foundation (Grant No. 1413/17). Rani Moran is a member of the Max Planck Centre for Computational Psychiatry and Ageing research at UCL, which is funded by the Max Planck Society, Munich, Germany, URL: https://www.mpg.de/en, Grant number: 647070403019. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript. We thank T. Sharot, B. Blain, I. Cogliati Dezza, L. Globig, C. Kelly, V. Vellani, S. Zheng, D. Lee and N. Nachman for critical reading of the manuscript and helpful comments.

**Author contributions:**

M.G and M.U. developed the study concept. M.G., R.M. and M.U. designed the experiments. M.G. and R.M. performed the experiments. M.G. analyzed the data and carried out the computational modeling. M.G., R.M. and M.U. wrote the paper and contributed to data discussion and interpretation at all stages.

**Competing interests**

The authors declare no competing interests.

## References


1. Bogacz, R., Brown, E., Moehlis, J., Holmes, P. & Cohen, J. D. The physics of optimal decision making: A formal analysis of models of performance in two-alternative forced-choice tasks. *Psychol. Rev.* **113**, 700–765 (2006).
2. Gold, J. I. & Shadlen, M. N. Banburismus and the brain: Decoding the relationship between sensory stimuli, decisions, and reward. *Neuron* **36**, 299–308 (2002).
3. Moran, R. Optimal decision making in heterogeneous and biased environments. *Psychon. Bull. Rev.* **22**, 38–53 (2015).
4. Wald, A. Foundations of a General Theory of Sequential Decision Functions. *Econometrica* **15**, 279 (1947).
5. Forstmann, B. U., Ratcliff, R. & Wagenmakers, E. J. Sequential sampling models in cognitive neuroscience: Advantages, applications, and extensions. *Annu. Rev. Psychol.* **67**, 641–666 (2016).
6. Brown, S. D. & Heathcote, A. The simplest complete model of choice response time: Linear ballistic accumulation. *Cogn. Psychol.* **57**, 153–178 (2008).
7. Gold, J. I. & Shadlen, M. N. Neural computations that underlie decisions about sensory stimuli. *Trends Cogn. Sci.* **5**, 10–16 (2001).
8. Ratcliff, R. & McKoon, G. The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Comput.* **20**, 873–922 (2008).
9. Teodorescu, A. R. & Usher, M. Disentangling decision models: From independence to competition. *Psychol. Rev.* **120**, 1–38 (2013).
10. Usher, M. & McClelland, J. L. The time course of perceptual choice: The leaky, competing accumulator model. *Psychol. Rev.* **108**, 550–592 (2001).
11. Vickers, D. Evidence for an accumulator model of psychophysical discrimination. *Ergonomics* **13**, 37–58 (1970).
12. Wickelgren, W. A. Speed-accuracy tradeoff and information processing dynamics. *Acta Psychol. (Amst).* **41**, 67–85 (1977).
13. Gold, J. I. & Shadlen, M. N. *The neural basis of decision making*. *Annual Review of Neuroscience* vol. 30 (2007).
14. Mulder, M. J., van Maanen, L. & Forstmann, B. U. Perceptual decision

neurosciences - a model-based review. *Neuroscience* **277**, 872–884 (2014).

15. Latimer, K. W., Yates, J. L., Meister, M. L. R., Huk, A. C. & Pillow, J. W. Single-trial spike trains in parietal cortex reveal discrete steps during decision-making. *Science (80-. ).* **349**, 184–187 (2015).
16. Watson, A. B. Probability summation over time. *Vision Res.* **19**, 515–522 (1979).
17. Stine, G. M., Zylberberg, A., Ditterich, J. & Shadlen, M. N. *Differentiating between integration and non-integration strategies in perceptual decision making*. *eLife* vol. 9 (2020).
18. Balci, F. *et al.* Acquisition of decision making criteria: Reward rate ultimately beats accuracy. *Attention, Perception, Psychophys.* **73**, 640–657 (2011).
19. Bogacz, R., Hu, P. T., Holmes, P. J. & Cohen, J. D. Do humans produce the speed-accuracy trade-off that maximizes reward rate? *Q. J. Experiment Psychol.* **63**, 863–891 (2010).
20. Palestro, J. J., Weichart, E., Sederberg, P. B. & Turner, B. M. Some task demands induce collapsing bounds: Evidence from a behavioral analysis. *Psychon. Bull. Rev.* **25**, 1225–1248 (2018).
21. Ditterich, J. Evidence for time-variant decision making. *Eur. J. Neurosci.* **24**, 3628–3641 (2006).
22. Voskuilen, C., Ratcliff, R. & Smith, P. L. Comparing fixed and collapsing boundary versions of the diffusion model. *J. Math. Psychol.* **73**, 59–79 (2016).
23. Glickman, M., Tsetsos, K. & Usher, M. Attentional Selection Mediates Framing and Risk-Bias Effects. *Psychol. Sci.* **29**, 2010–2019 (2018).
24. Glickman, M. *et al.* The formation of preference in risky choice. *PLoS Comput. Biol.* **15**, e1007201 (2019).
25. Gluth, S., Kern, N., Kortmann, M. & Vitali, C. L. Value-based attention but not divisive normalization influences decisions with multiple alternatives. *Nat. Hum. Behav.* **4**, 634–645 (2020).
26. Krajbich, I., Armel, C. & Rangel, A. Visual fixations and the computation and comparison of value in simple choice. *Nat. Neurosci.* **13**, 1292–1298 (2010).
27. Brehm, J. W. Postdecision changes in the desirability of alternatives. *J. Abnorm. Soc. Psychol.* **52**, 384–389 (1956).

28. Bronfman, Z. Z. *et al.* Decisions reduce sensitivity to subsequent information. *Proc. R. Soc. B Biol. Sci.* **282**, 20150228 (2015).
29. Kappes, A., Harvey, A. H., Lohrenz, T., Montague, P. R. & Sharot, T. Confirmation bias in the utilization of others' opinion strength. *Nat. Neurosci.* **23**, 130–137 (2020).
30. Rollwage, M. *et al.* Confidence drives a neural confirmation bias. *Nat. Commun.* **11**, 1 (2020).
31. Talluri, B. C., Urai, A. E., Tsetsos, K., Usher, M. & Donner, T. H. Confirmation Bias through Selective Overweighting of Choice-Consistent Evidence. *Curr. Biol.* **28**, 3128-3135.e8 (2018).
32. Spitzer, B., Waschke, L. & Summerfield, C. Selective overweighting of larger magnitudes during noisy numerical comparison. *Nat. Hum. Behav.* **1**, 1 (2017).
33. Tsetsos, K. *et al. Economic irrationality is optimal during noisy decision making*. *Proceedings of the National Academy of Sciences of the United States of America* vol. 113 (2016).
34. Glickman, M. & Usher, M. Integration to boundary in decisions between numerical sequences. *Cognition* **193**, 104022 (2019).
35. Thura, D. & Cisek, P. *Modulation of premotor and primary motor cortical activity during volitional adjustments of speed-accuracy trade-offs*. *Journal of Neuroscience* vol. 36 (2016).
36. Thura, D. & Cisek, P. The Basal Ganglia Do Not Select Reach Targets but Control the Urgency of Commitment. *Neuron* **95**, 1160-1170.e5 (2017).
37. van Maanen, L., Fontanesi, L., Hawkins, G. E. & Forstmann, B. U. *Striatal activation reflects urgency in perceptual decision making*. *NeuroImage* vol. 139 (Academic Press, 2016).
38. Hawkins, G. E., Forstmann, B. U., Wagenmakers, E. J., Ratcliff, R. & Brown, S. D. Revisiting the evidence for collapsing boundaries and urgency signals in perceptual decision-making. *J. Neurosci.* **35**, 2476–2484 (2015).
39. FISHER, R. A. the Use of Multiple Measurements in Taxonomic Problems. *Ann. Eugen.* **7**, 179–188 (1936).
40. McLachlan, G. J. *Discriminant Analysis and Statistical Pattern Recognition*. (John

Wiley & Sons, Inc., 2005).

41. Dotan, D., Meyniel, F. & Dehaene, S. On-line confidence monitoring during decision making. *Cognition* **171**, 112–121 (2018).
42. Usher, M., Tsetsos, K., Glickman, M. & Chater, N. Selective Integration: An Attentional Theory of Choice Biases and Adaptive Choice. *Curr. Dir. Psychol. Sci.* **28**, 552–559 (2019).
43. Zeigenfuse, M. D., Pleskac, T. J. & Liu, T. Rapid decisions from experience. *Cognition* **131**, 181–194 (2014).
44. Ossmy, O. *et al.* The timescale of perceptual evidence integration can be adapted to the environment. *Curr. Biol.* **23**, 981–986 (2013).
45. Roe, R. M., Busemeyer, J. R. & Townsend, J. T. Multialternative decision field theory: A dynamic connectionist model of decision making. *Psychol. Rev.* **108**, 370–392 (2001).
46. Balsdon, T., Wyart, V. & Mamassian, P. Confidence controls perceptual evidence accumulation. *Nat. Commun.* **11**, 1 (2020).
47. Eisen-Enosh, A., Farah, N., Burgansky-Eliash, Z., Polat, U. & Mandel, Y. Evaluation of Critical Flicker-Fusion Frequency Measurement Methods for the Investigation of Visual Temporal Resolution. *Sci. Rep.* **7**, 1 (2017).
48. Ludwig, C. J. H., Gilchrist, I. D., McSorley, E. & Baddeley, R. J. The temporal impulse response underlying saccadic decisions. *J. Neurosci.* **25**, 9907–9912 (2005).
49. Britten, K. H., Shadlen, M. N., Newsome, W. T. & Movshon, J. A. The analysis of visual motion: A comparison of neuronal and psychophysical performance. *J. Neurosci.* **12**, 4745–4765 (1992).
50. Shadlen, M. N. & Newsome, W. T. Neural basis of a perceptual decision in the parietal cortex (area LIP) of the rhesus monkey. *J. Neurophysiol.* **86**, 1916–1936 (2001).
51. Yang, T. & Shadlen, M. N. Probabilistic reasoning by neurons. *Nature* **447**, 1075–1080 (2007).
52. Luyckx, F., Spitzer, B., Blangero, A., Tsetsos, K. & Summerfield, C. Selective Integration during Sequential Sampling in Posterior Neural Signals. *Cereb. Cortex*

**30**, 4454–4464 (2020).

53. Tsetsos, K., Chater, N. & Usher, M. *Salience driven value integration explains decision biases and preference reversal*. *Proceedings of the National Academy of Sciences of the United States of America* vol. 109 (2012).
54. Abrahamyan, A., Silva, L. L., Dakin, S. C., Carandini, M. & Gardner, J. L. Adaptable history biases in human perceptual decisions. *Proc. Natl. Acad. Sci. U. S. A.* **113**, E3548–E3557 (2016).
55. Braun, A., Urai, A. E. & Donner, T. H. Adaptive history biases result from confidence-weighted accumulation of past choices. *J. Neurosci.* **38**, 2418–2429 (2018).
56. Urai, A. E., De Gee, J. W., Tsetsos, K. & Donner, T. H. *Choice history biases subsequent evidence accumulation*. *eLife* vol. 8 (2019).
57. Rollwage, M., Dolan, R. J. & Fleming, S. M. Metacognitive Failure as a Feature of Those Holding Radical Beliefs. *Curr. Biol.* **28**, 4014-4021.e8 (2018).
58. Jazayeri, M. & Movshon, J. A. A new perceptual illusion reveals mechanisms of sensory decoding. *Nature* **446**, 912–915 (2007).
59. Luu, L. & Stocker, A. A. *Post-decision biases reveal a self-consistency principle in perceptual inference*. *eLife* vol. 7 (2018).
60. Stocker, A. A. & Simoncelli, E. P. A Bayesian model of conditioned perception. *Adv. Neural Inf. Process. Syst. 20* 1409–1416 (2007).
61. Cheadle, S. *et al.* Adaptive gain control during human perceptual choice. *Neuron* **81**, 1429–1441 (2014).
62. Patai, Z. E. *et al.* Conflict detection in a sequential decision task is associated with increased cortico-subthalamic coherence and prolonged subthalamic oscillatory response in the beta band. *bioRxiv* 453 (2020) doi:10.1101/2020.06.09.141713.
63. Kruglanski, A. W. & Webster, D. M. Motivated Closing of the Mind: 'Seizing' and 'Freezing'. *Psychol. Rev.* **103**, 263–283 (1996).
64. Schulz, L., Rollwage, M., Dolan, R. J. & Fleming, S. M. Dogmatism manifests in lowered information search under uncertainty. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 31527–31534 (2020).
65. Cavedini, P., Gorini, A. & Bellodi, L. Understanding obsessive-compulsive

disorder: Focus on decision making. *Neuropsychol. Rev.* **16**, 3–15 (2006).

66. Peirce, J. *et al.* PsychoPy2: Experiments in behavior made easy. *Behav. Res. Methods* **51**, 195–203 (2019).
67. Teodorescu, A. R., Moran, R. & Usher, M. Absolutely relative or relatively absolute: violations of value invariance in human decision making. *Psychon. Bull. Rev.* **23**, 22–38 (2016).
68. Luce, R. D. (Robert D. *Individual choice behavior : a theoretical analysis / R. Duncan Luce*. (Wiley ; Chapman & Hall, 1959).
69. Tavares, G., Perona, P. & Rangel, A. The attentional Drift Diffusion Model of simple perceptual decision-making. *Front. Neurosci.* **11**, 468 (2017).
70. Nelder, J. A. & Mead, R. A Simplex Method for Function Minimization. *Comput. J.* **7**, 308–313 (1965).
71. White, C. N., Servant, M. & Logan, G. D. Testing the validity of conflict drift-diffusion models for use in estimating cognitive processes: A parameter-recovery study. *Psychon. Bull. Rev.* **25**, 286–301 (2018).

# Evidence integration and decision-confidence are modulated by stimulus consistency

Moshe Glickman, Rani Moran and Marius Usher

## Supplementary Information

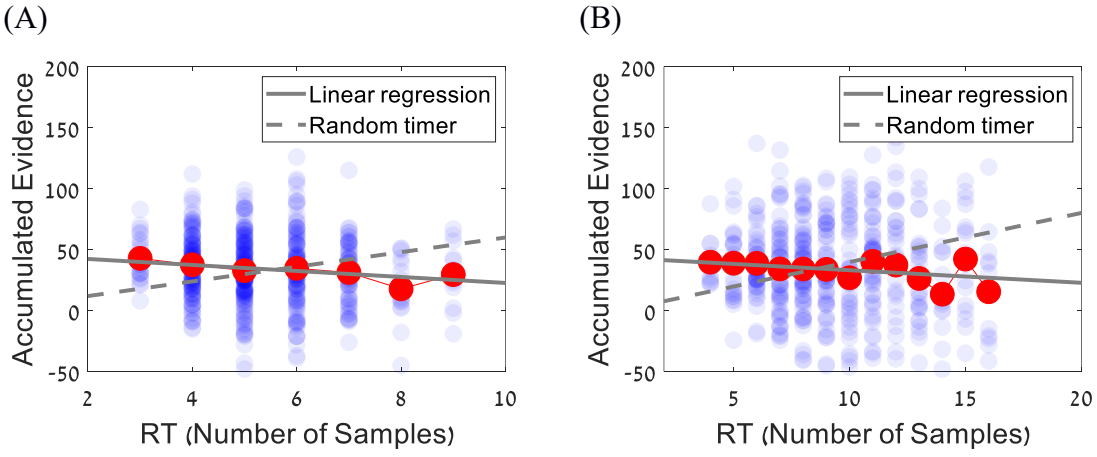
# Supplementary Results

## Behavioral signature of integration to boundary

We compared integration and non-integration models by using the method described in ref. [1]. Each model was simulated 100,000 times and the mean accumulated evidence was plotted as a function of RT (see black, green and magenta dashed lines in Fig. 1C-D). Next, for each participant in Experiment 1 & 2, the accumulated evidence excluding internal noise (both of correct and incorrect responses) was plotted as a function of response-time (see Supplementary Figure 1A-B). A mixed-effect linear regression models were fitted to the data of Experiment 1 & 2, with RT and participants serving as random effects and with a free covariance matrix (see thin gray and blue lines in Fig. 1C-D). The model in Wilkinson notation was:

$$Accumulated\ evidence\ at\ response \sim RT\ +\ (RT|subject)$$

The group level slope in both Experiment 1 & 2 showed a decreasing pattern (Fig. 1C: $b$ = -2.86, $t_{\text{against 0}}$ = -7.08, $p$ < .001, 95% CI -3.65 to -2.07 and Fig. 1D: $b$ = -0.86, $t_{\text{against 0}}$ = -3.18, $p$ = .001, 95% CI -1.40 to -0.33), which is the behavioral signature of integration to a collapsing boundary.




***Supplementary Figure 1.*** *Accumulated evidence as a function of RT. (A) Scatter-plot showing the accumulated evidence in individual trials (blue circles) as a function of decision time of a representative participant in Experiment 1 (B) Same as (A) but for a representative participant in Experiment 2. Red circles correspond to the average accumulated evidence at each RT, the gray solid line corresponds to linear regression fitted to the data, and the dashed gray line corresponds to the prediction of the random timer model (see Computational methods).*

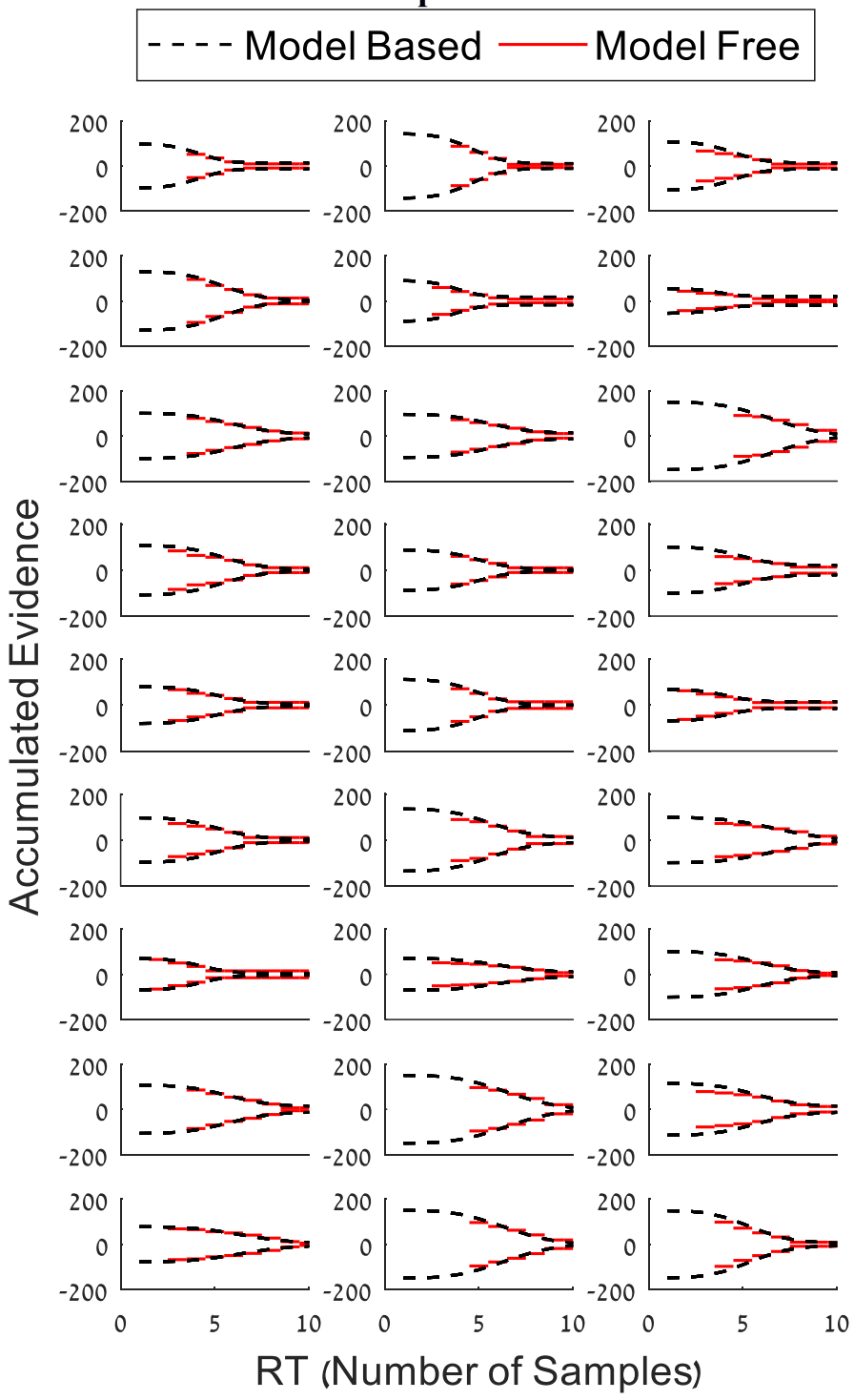
## Regression analysis - Largest temporal cluster of the evidence

The choice-accuracy, RT and confidence in Experiment 1 & 2 were analyzed using mixed model regression analyses (logistic for accuracy and linear for RT and confidence), using the accumulated evidence and the normalized stimulus-consistency as fixed factors and participants as random intercepts. The analysis was similar to one reported in Table 1 (see also Computational models), except that here we used the more complex measure of stimulus-consistency – larger temporal cluster of the evidence (LTC), in which we computed for each trial the largest temporal cluster that goes with the evidence and divide it by the number of samples (e.g., if the evidence are: 4, -5, 3, 4, 5, 2, -2, -6, then the LTC measure will be equal to: $\frac{4}{8} = 0.5$). The results (Supplementary Table 1), are fully consistent with the ones obtained using the simpler consistency measure (Difference in Evidence-Directions), and show that LTC enhances accuracy and confidence and reduces RT, even whilst controlling for the accumulated evidence.

***Supplementary Table 1.*** *Beta coefficients for predicting Accuracy, RT and confidence in Experiment 1 & 2.*

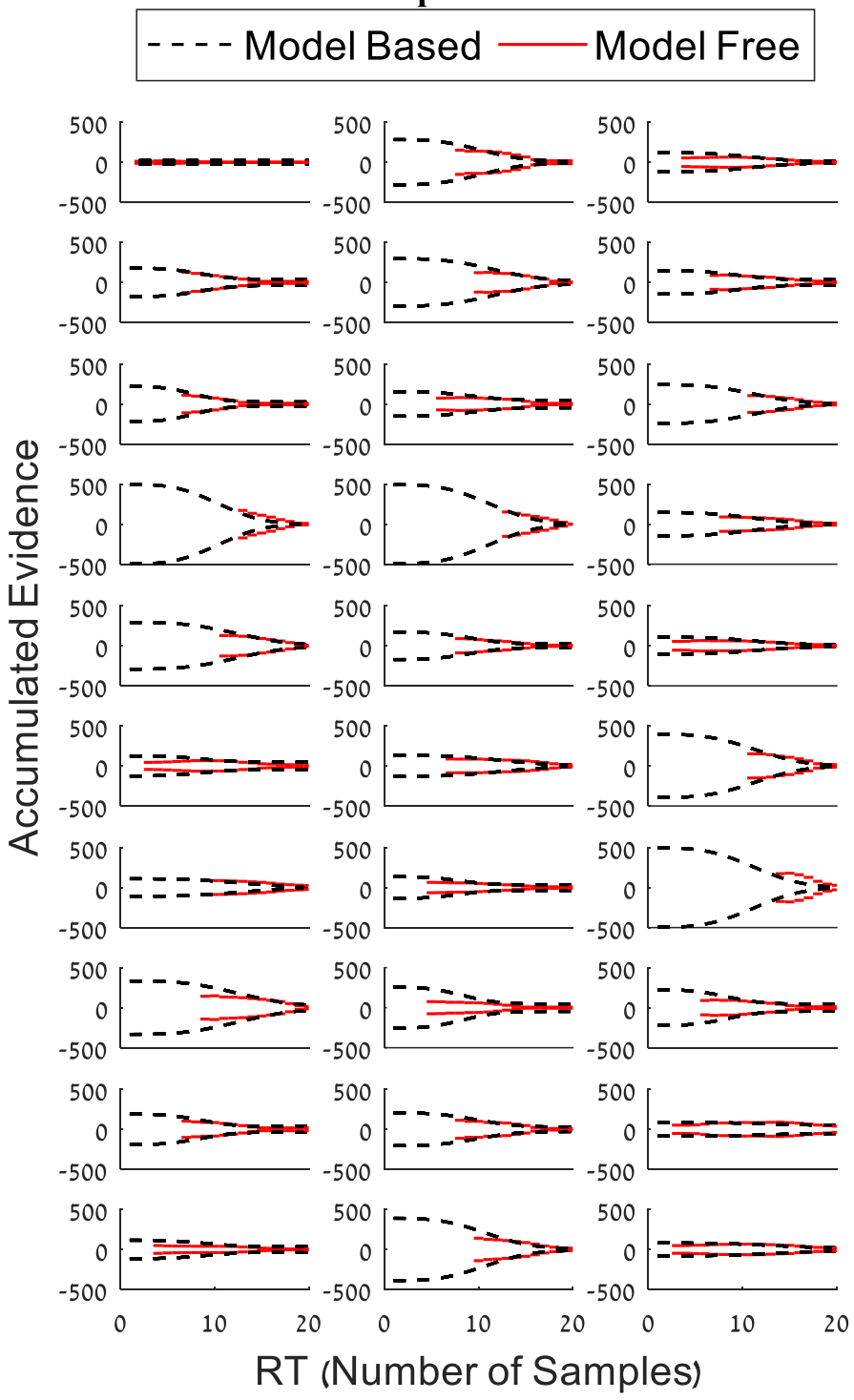
| | *β (S.E.)* | *t* | *p* | 95% *CI* |
|---|---|---|---|---|
| Experiment 1 | | | | |
| Accuracy | | | | |
| Evidence | 1.80 (0.06) | 32.17 | <.001 | [1.69, 1.91] |
| Stimulus Consistency (LTC) | 0.20 (0.04) | 4.85 | <.001 | [0.12, 0.28] |
| RT | | | | |
| Evidence | -0.29 (0.008) | -35.39 | <.001 | [-0.30, -0.27] |
| Stimulus Consistency (LTC) | -0.14 (0.008) | -16.97 | <.001 | [-0.15, -0.12] |
| Experiment 2 | | | | |
| Accuracy | | | | |
| Evidence | 0.53 (0.03) | 17.58 | <.001 | [0.47, 0.59] |
| Stimulus Consistency (LTC) | 0.97 (0.04) | 27.43 | <.001 | [0.90, 1.04] |
| RT | | | | |
| Evidence | -0.19 (0.007) | -28.12 | <.001 | [-0.20,-0.17] |
| Stimulus Consistency (LTC) | -0.16 (0.007) | -24.26 | <.001 | [-0.15,-0.12] |
| Confidence | | | | |
| Evidence | 0.18 (0.009) | 20.96 | <.001 | [0.16, 0.20] |
| Stimulus Consistency (LTC) | 0.10 (0.009) | 10.99 | <.001 | [0.08, 0.11] |

**Model based vs. Model free - Experiment 1**




***Supplementary Figure 2.*** *Decision Classification Boundaries (DCB) of the participants in Experiment 1. Black dashed lines correspond to the model-based boundary, and the red ones correspond to the model-free DCB.*

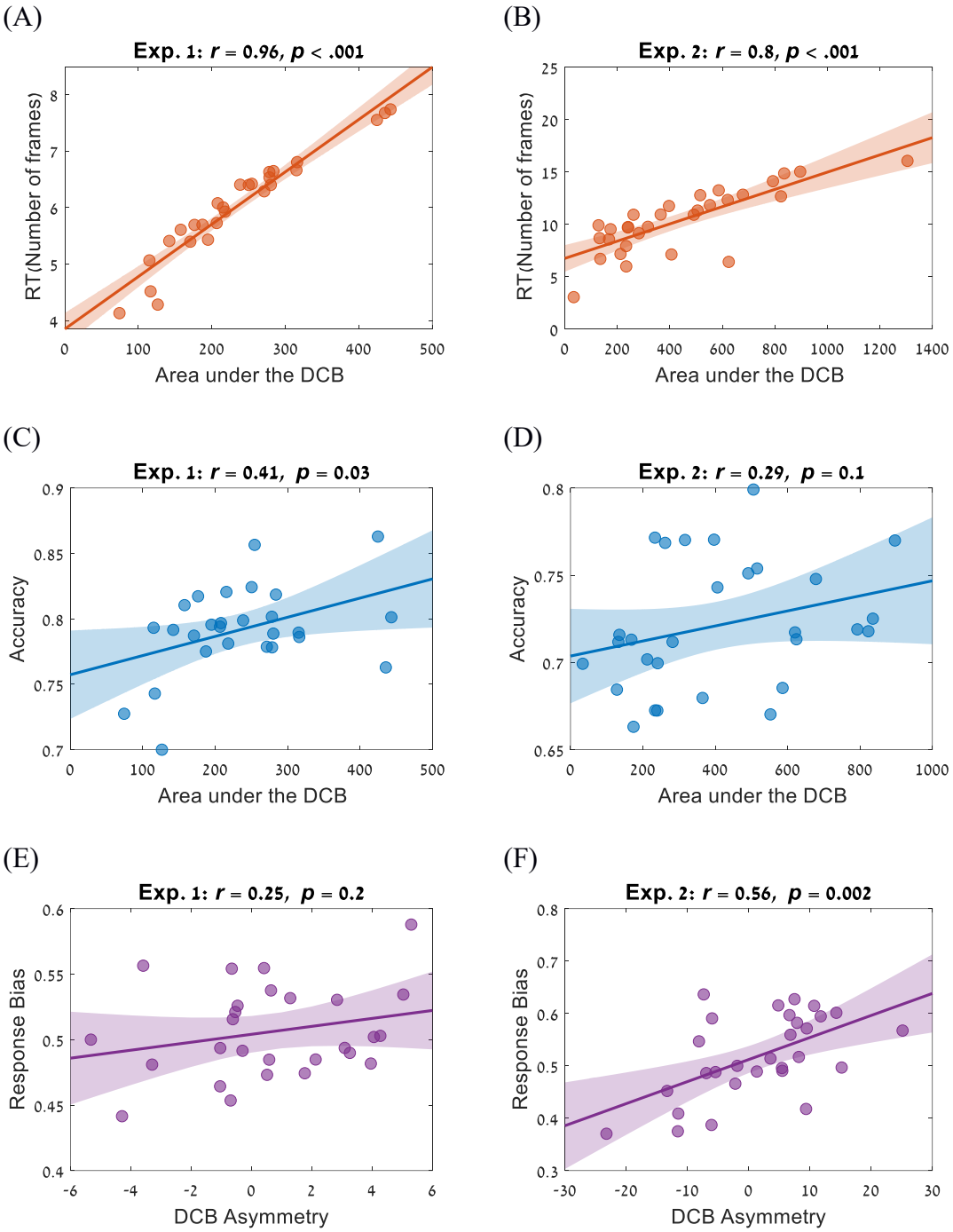
**Model based vs. Model free - Experiment 2**




***Supplementary Figure 3.*** *Decision Classification Boundaries (DCB) of the participants in Experiment 2. Black dashed lines correspond to the model-based boundary, and the red ones correspond to the model-free DCB.*

## Linking the DCB to behavioral measures

The link between the DCB and several behavioral measures, such as response-times, accuracy and response bias, was examined using Pearson correlations. The correlation between the area under the DCB and the mean RT across participants was high and significant in both experiments, Experiment 1: $r = .96$, $p < .001$, Experiment 2: $r = .8$, $p < .001$ (Supplementary Figure 4A-B). There was also a correlation between the area under the DCB and the mean accuracy, however, it was weaker and less consistent – Experiment 1: $r = .41$, $p = .03$ (Supplementary Figure 4C) and Experiment 2: $r = .29$, $p = .1$ (Supplementary Figure 4D). In the analysis of Experiment 2, one of the participant was considered an outlier (z-scored Mahalanobis distance > 4) and was excluded from the analysis. Finally, we examined the correlation between the DCB and response bias. To this end, we computed the asymmetry between the area under the boundaries (that is, difference between the area under the left and right boundaries) and correlated it with the probability to choose the left alternative over the right one (that is, response bias). The correlation between this two measures did not reach a statistical significance in Experiment 1: $r = .25$, $p = .2$ (Supplementary Figure 4E), but was significant in Experiment 2: $r = .55$, $p = .002$ (Supplementary Figure 4F).




***Supplementary Figure 4.*** *The links between the DCB and behavioral measures. (A-B) Correlations between the area under the DCB and the mean response times across participants in (A) Experiment 1 (r = .96, p < .001) and (B) Experiment 2 (r = .8, p < .001). (C-D) Correlations between the area under the DCB and the mean accuracy across participants in (C) Experiment 1 (r = .41, p = .03) and (D) Experiment 2 (r = .29, p = .1). (E-F) Correlation between the asymmetry of the area under the boundary and response bias in (E) Experiment 1 (r = .25, p = .2) and (F) Experiment 2 (r = .56, p = .002). Solid lines correspond to linear least-squares fits, and the shaded areas to 95% confidence intervals.*

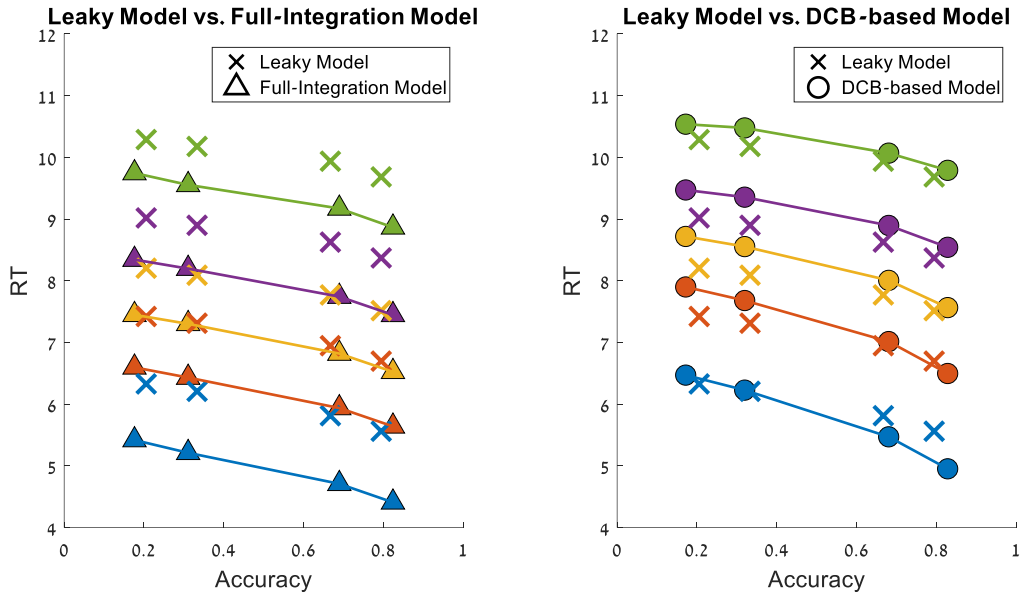
## DCB for biased evidence – bias free approximate model

The DCB accurately extracts the decision-boundary when the evidence is integrated without loss or distortion (that is, full integration model). However, if this is not the case, the DCB does not recover the decision-boundary, but rather modulates the classification curve so as to produce a bias-free model of evidence integration which approximates the original model. To illustrate this point, we simulated three models: i) Leaky-integration model:

$$DV(t) = (1-\varphi)\cdot DV(t-1) + \mu(t) + \varepsilon(t),\ \varepsilon \sim N(0,\sigma^2) \qquad \text{Eq. (S1)}$$

where *DV(t)* is the accumulated differences between the sequences at time *t*, $\mu(t)$ is the difference between the samples at time *t*, $\varepsilon(t)$ is a temporally-independent random internal Gaussian noise, and $\varphi$ is a leak parameter which controls the extent to which earlier values are given less weight[2] and was set to 0.3 in this simulation. ii) Full integration model (that is, the leak was set to 0) with the same boundary parameters as of the previous leaky-integration model, and iii) Full integration model with a boundary that was extracted by applying the DCB on the data of the leaky-integration model (model i). That is, the data of this model was generated by the extracting the DCB of the leaky–integration model first, and then a non-leaky model was simulated using the DCB values as decision boundary (all other parameters except of leak are the same as in the original model). Each model was simulated 10,000 times, using two difficulty condition: difficult trials, in which the values were sampled from $X \sim N(52, 10^2)$ and $Y \sim N(48, 10^2)$, and easy trials in which the values were sampled from $X \sim N(52, 10^2)$ and $Y \sim N(44, 10^2)$. In all models we assume integration to a collapsing-boundary characterized by intercept (a=150), shape (k=3), scale (λ=5) and asymptote (a'=10) parameters[3] with noise level of σ = 15. Supplementary Figure 5A shows the Quantile Probability Plots (QPP) of the leaky-integration model and the full integration model (without leak) with the same boundary parameters. As expected, the QPPs of these models show large deviations from each other. By contrast, as shown in Supplementary Figure 5B, the QPPs of the leaky-integration model and the (full integration) DCB-based model are much more similar, indicating that the DCB compensate for the distorted integration by modulation of the classification curve to maintain similar choice and response-times patterns as of the leaky-integration model.

(A) (B)




***Supplementary Figure 5.*** *Quantile Probability Plots (QPP) of the leaky integration model ('X' symbol), full integration model ('△' symbol) and DCB-based model ('○' symbol). The blue, orange, yellow, purple and green lines correspond to the 0.1, 0.3, 0.5, 0.7 and 0.9 quantiles, respectively. (A) Comparison between the QPPs of the leaky and non-leaky integration models. (B) Comparison between the QPPs of the leaky and DCB-based integration models.*

## DCB for biased evidence – Estimation of the boundary

We have showed that the DCB is able to accurately estimate the decision boundary when the integration process is not biased. Here, we demonstrate how the DCB method can be applied to estimate the boundary even in case of biased integration. To this end, a candidate biased integration model (e.g., stimulus-consistency bias) is first selected. Then, instead of extracting the DCB using the actual evidence, the DCB is applied on the evidence biased by the selected model. This procedure is repeated for different levels of the bias-parameter of the model (e.g., the different levels of the stimulus-consistency bias parameter). For each level, the corresponding DCB and a classification performance metric (e.g., accuracy) are computed. This results in a family of DCB curves – one for each value of the bias parameter. The bias-parameter and corresponding DCB which maximizes the performance metric comprise the estimate. Note that this method is based on a mixture of parametric (the proposed biased integration mechanism) and non-parametric methods (the DCB, which does not assume any functional form of the boundary). Thus, this method allows to extract the boundary without making any functional assumptions using the DCB, even the when evidence integration process is biased.

To compute the DCB performance measure, the confusion matrix of the DCB response classification (aggregated across all frames) should be first defined. This is a 3X3 matrix for the classes: a) continue sampling, b) choosing alternative A, and c) choosing alternative B. The rows of the matrix correspond to the actual responses and the columns corresponds to the responses predicted by the DCB. The DCB classification performance measure can be obtained using standard metrics such as accuracy, F1-score[4] or Matthews correlation coefficient[5] (MCC). Note that instead of maximizes a performance metric, one can also minimize an error measure (e.g., the overlap between the 3 classes of distributions; see Fig. 2).

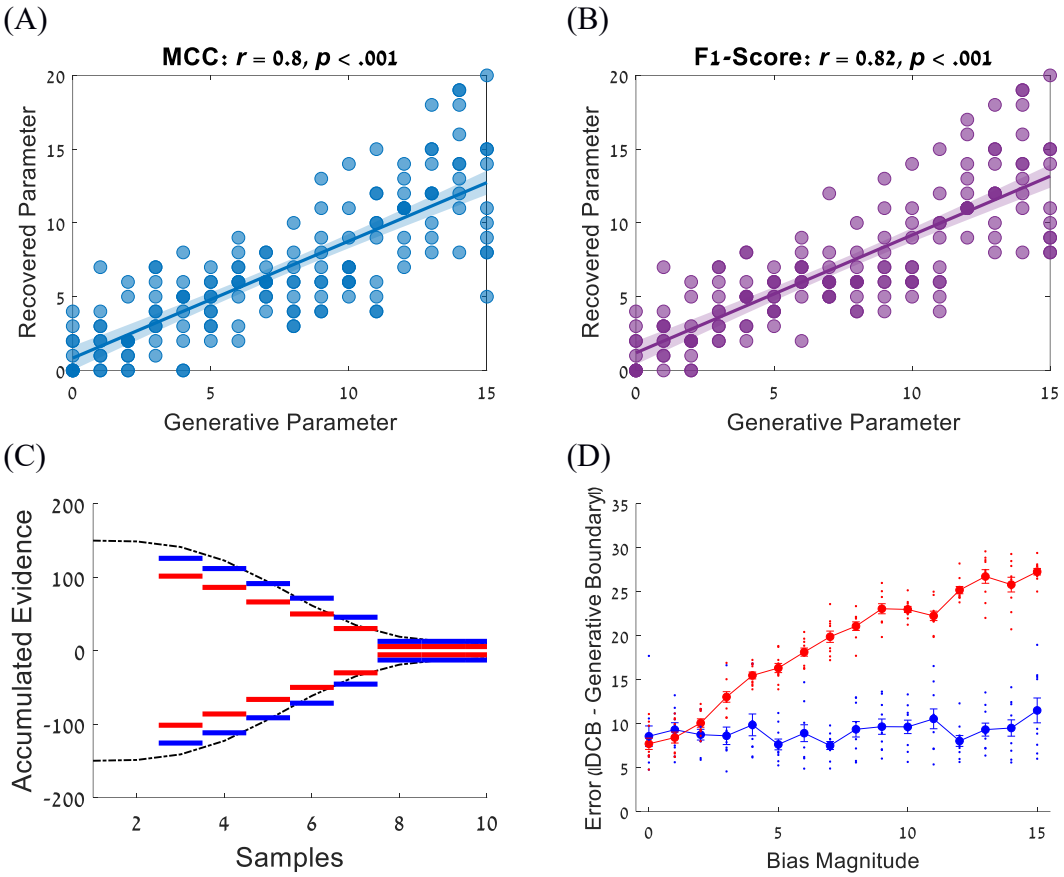
***Supplementary Table 2.*** *The confusion matrix of the DCB.*

| | | Predicted response | | |
|---|---|---|---|---|
| | | Continue Sampling | Choosing A | Choosing B |
| Actual Response | Continue Sampling | Correct Classification | Incorrect Classification | Incorrect Classification |
| | Choosing A | Incorrect Classification | Correct Classification | Incorrect Classification |
| | Choosing B | Incorrect Classification | Incorrect Classification | Correct Classification |

To validate this method, we simulated data using different stimulus-consistency bias parameters and examine the ability of the method to recover them. For each bias parameter, the model is simulated 1,000 times using collapsing-boundary characterized by four parameters: intercept, shape, scale and asymptote parameters (a=150, k=3, λ=5, a'=10; see ref. [3]) and noise level of σ = 15. The evidence is sampled from $X \sim N(52, 10^2)$ and $Y \sim N(46, 10^2)$. The consistency bias-parameters with which the data was generated ranged from 0 to 15. For each parameter, we used a grid search over values between 0 and 20 with increments of 1.

Supplementary Figure 6A-B show the generative stimulus-consistency bias-parameters plotted against the recovered parameters. The parameters were recovered using Matthews correlation coefficient (Supplementary Figure 6A) and F1-score (Supplementary Figure 6B) metrics, which are better suited for imbalanced data. Supplementary Figure 6C shows an illustrative simulation in which the stimulus-consistency bias parameter was set to 6. As can be seen, the DCB assuming a biased evidence integration process (in blue) captured well the generative boundary with which the data was simulated (black lines). By contrast, a DCB assuming a bias of 0 (that is, unbiased integration; red lines) underestimates the generative boundary. Note that as shown in Supplementary Figure 5, a DCB assuming a bias of 0 compensates for the biased integration process by modulation of the classification curves, in order to maintain a similar choice and response-times patterns. In this case, as the stimulus-consistency bias increases the evidence, the DCB compensates by underestimating the generative boundary. Supplementary Figure 6D shows the boundary reconstruction error as a function of bias-parameter. Reconstruction error was quantified as the mean absolute deviation between the generative boundary and the DCB (that is, $\text{Error} = < \Sigma|\text{Generative boundary}_i - \text{Estimated DCB}_i| >$). It was calculated both for the DCB

that minimizes the error (in blue) as well as for the DCB assuming unbiased integration (in red). As shown in Supplementary Figure 6D, the reconstruction error of the former is lower and does not depended on the magnitude of the bias (blue line), whereas the error of the latter is higher and increases with the magnitude of the bias-parameter (red line).

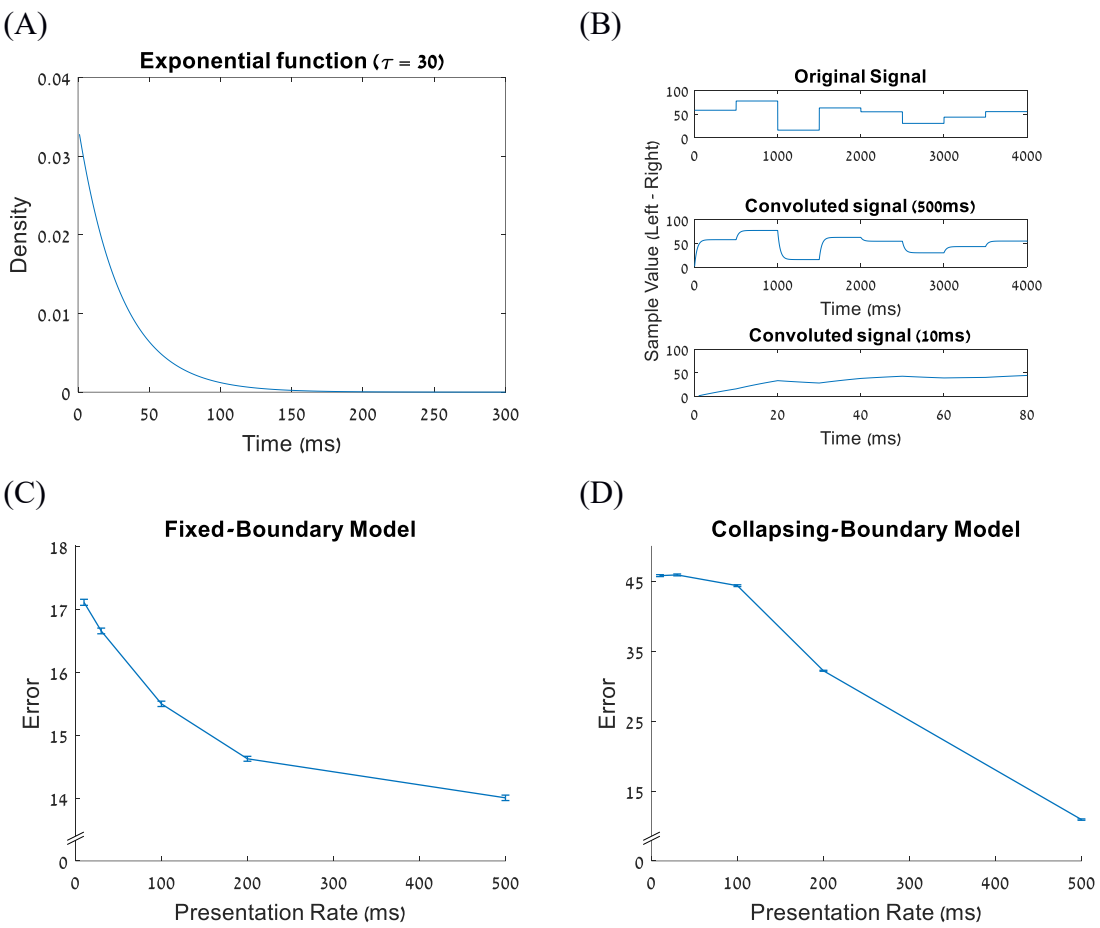



***Supplementary Figure 6.*** *DCB for biased integration process. (A-B). The generative stimulus-consistency bias parameters plotted against the recovered parameters using the (A) MCC* ($r = .8$, $p < .001$) *and (B) F1-score* ($r = .82$, $p < .001$) *as a goodness of fit measures. Solid lines correspond to linear least-squares fits, and the shaded areas to 95% confidence intervals. (C) Simulation illustrates that applying the DCB (blue line) whilst assuming a biased evidence integration process allows to recover the generative boundary with which the data was generated (black). The red line corresponds to a model that assumes an unbiased evidence integration and results in a poorer recovery of the generative boundary. (D) Reconstruction error as a function of bias-parameter (n = 160 simulations). Reconstruction error was quantified as the mean absolute deviation between the generative boundary and the DCB. It was calculated both for the DCB that maximizes performance measures (in blue) as well as for the DCB assuming unbiased integration (in red). As shown, the reconstruction error of the former is lower and does not depended on the magnitude of the bias, whereas the error of the latter is higher and increases with the magnitude of the bias-parameter. Red and blue dots correspond to the reconstruction error of each simulation. Data are presented as mean values ± standard error of the mean.*

## DCB error as a function of presentation rate

The ability of the DCB to reconstruct the decision boundary for presentation rates of 500ms/frame (Experiment 1) and 200ms/frame (Experiment 2) is described in the main text. Here, we theoretically examined (using simulations) the ability of the DCB to reconstruct the boundary also for faster presentation rates, which are closer to the visual integration time-constant (assumed here to correspond to a conservative estimation of $\tau$ = 30ms[6,7]). To address this question, we simulated two types of models: a fixed-boundary model (c=50) and a collapsing-boundary model characterized by intercept, shape, scale and asymptote parameters (a = 150, k = 3, $\lambda$ = 5, a' = 10; Hawkins et al., 2015). Each model was simulated 100 times with 10,000 trials in each simulation and using noise level of $\sigma$ = 15. The evidence in each simulation was sampled from $X \sim N(52, 10^2)$ and $Y \sim N(46, 10^2)$, as in our experiments. Critically, each simulation assumed five presentation rates of 10, 30, 100, 200 and 500ms/frame. For the collapsing boundary model, we assumed that the scale parameter was scaled with the presentation times. For example, if $\lambda$ = 5, then for presentation time of 200ms it would be equal to: $5 \cdot 500/200 = 7.5$ (the same pattern of results is obtained if we do not make this assumption, and assumed the same scale for all presentation rates). For each frame, we computed the convolution between a decaying exponential function with a time constant of $\tau$ = 30ms (Supplementary Figure 7A) and a step function with values $S_i = X_i - Y_i$ over intervals equal to the size of the frame durations (that is, the inverse of presentation rates; Supplementary Figure 7B, upper panel). The results of this convolution were averaged for each frame and fed as an input to the DCB. As shown in Supplementary Figure 7B (middle and lower panels), the output signal of the convolution was much more distorted for fast presentation rates compared to slow ones. To quantify the ability of the DCB to extract the generative boundary of each simulation, we computed the mean absolute difference between the DCB and the generative boundary (Error = $<\Sigma|\text{Generative boundary}_i - \text{Estimated DCB}_i|>$). Supplementary Figure 7C-D show the mean reconstruction error of the DCB as a function of presentation rate for the fixed-boundary and collapsing-boundary models, respectively. As can be seen, both panels show that the slower the presentation rate, the lower the error. These findings suggest that the DCB will be more reliable in paradigms in which the response time is not very fast (see also the discussion in the main text).

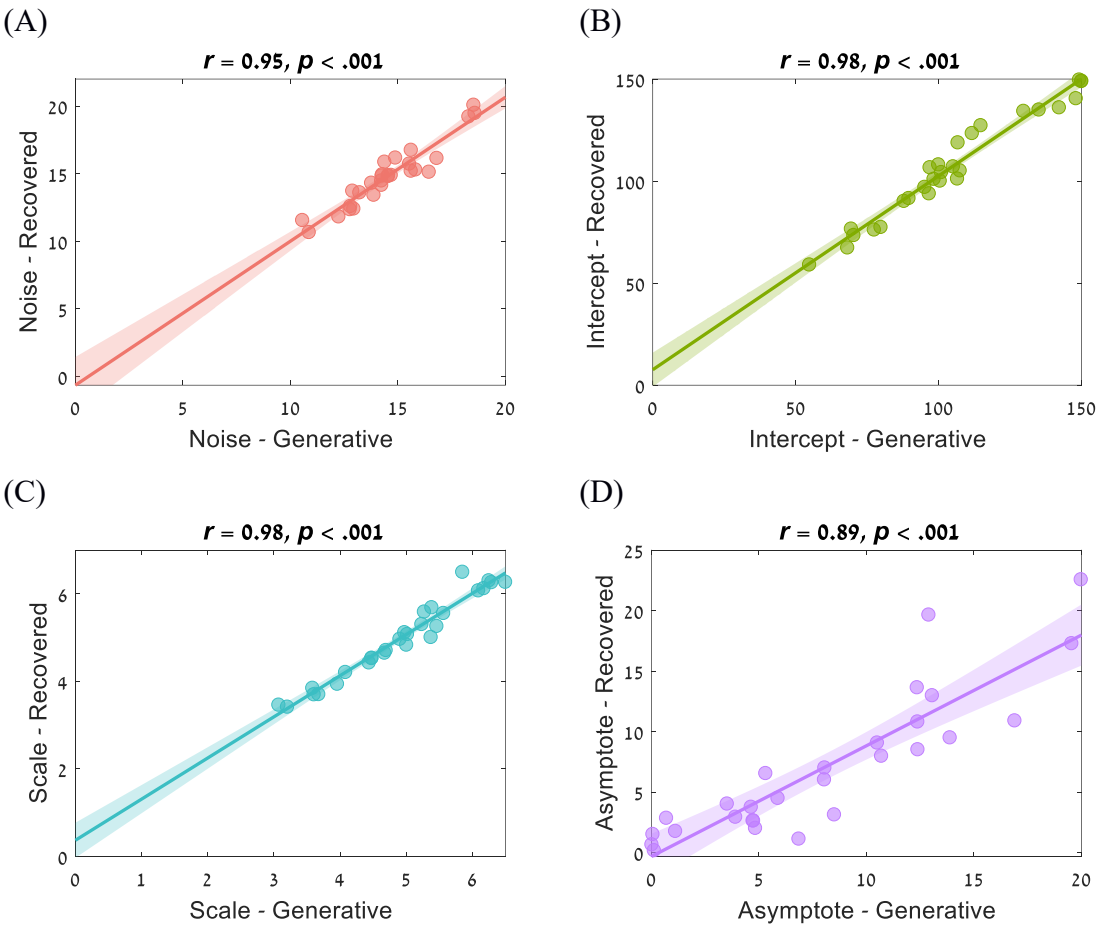
Note that in these simulations, we follow the normative task requirements of updating the evidence per sample (rather than per unit of time). This is because in our task, in which pairs of numbers from noisy distributions are presented, the likelihood changes with the samples rather than with the presentation time (that is, presenting a single sample twice as long is not equivalent with presenting two identical samples; we assume that the sampling noise is higher than the perceptual noise). We thus use the presentation frame as our updating simulation time steps. A similar normative algorithm is characteristic of tasks such as the weather prediction-task[8] or legal decision making. However, it is also possible that as the time frame is reduced, the estimation error (that is, the internal noise) is increased. Here, we focus on how the time frame of evidence integration limits the accuracy of the boundary estimation, as a result of the temporal fusion between the information from different frames which is caused by visual integration that takes place at the percept rather than decision level. For that reason, we neglect the dependency of the estimation error (at each sample) on the duration of the presentation rate. Thus, the results obtained from this simulation provide an upper limit for the DCB accuracy at fast presentation rates.



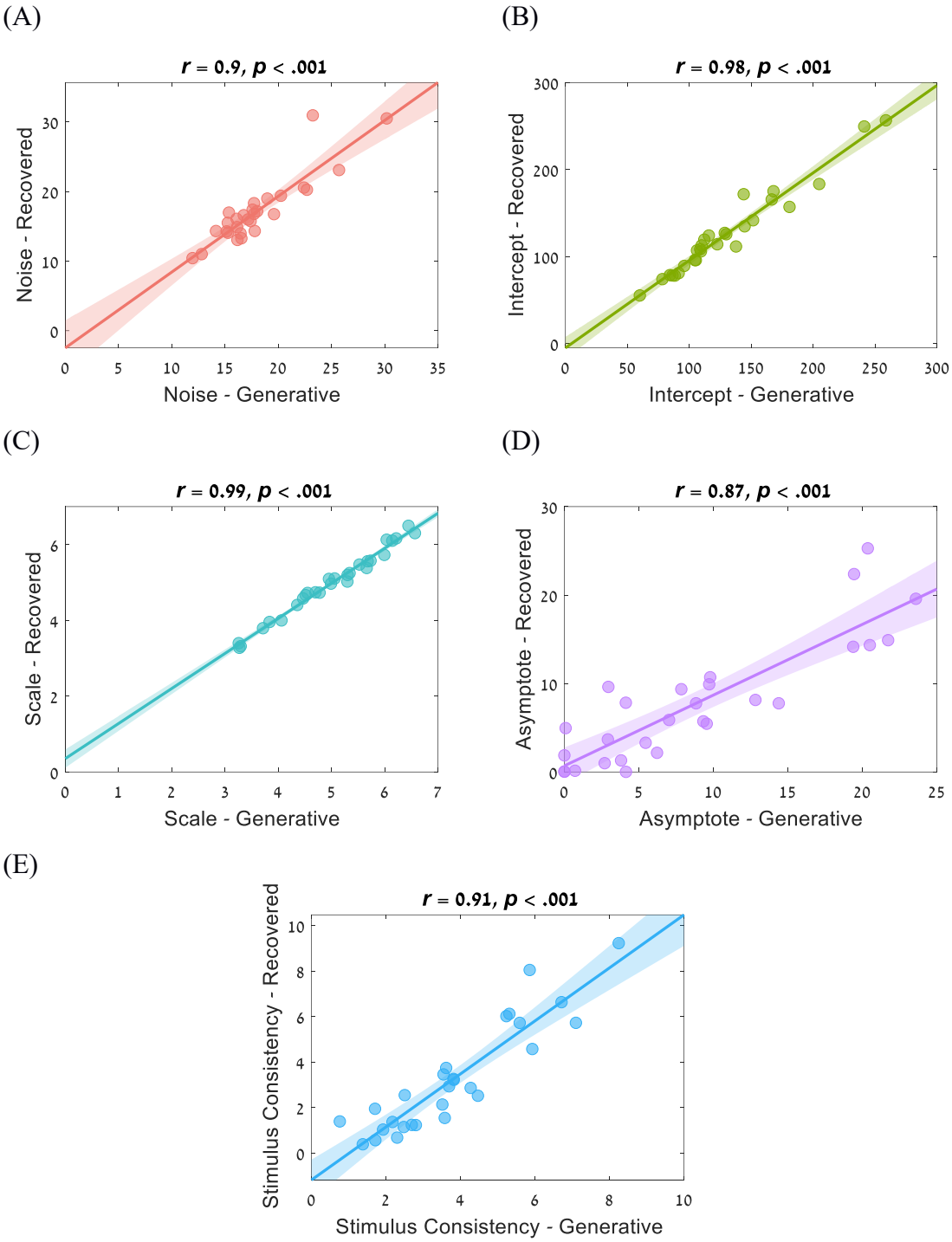

***Supplementary Figure 7.*** *(A) A decaying exponential function with a time constant of 30ms (τ), corresponding to the visual filter that transforms the signal on the retina to signals in the visual areas. (B) Upper panel: The sample values (Left sample – Right sample) which were presented in a representative trial as a function of time. Middle panel: the convolution between the decaying exponential function (τ=30ms) and the sampled values for presentation rate of 500ms/frame. Lower panel: same as the middle panel, but for presentation rate of 10ms/frame. (C-D) Both fixed and collapsing boundary model show that the mean reconstruction error of the DCB decreases with presentation rate (n = 100 simulations). Data are presented as mean values ± standard error of the mean.*

## Parameter recovery

Parameter recovery was performed for the full integration model and for the stimulus-consistency model (our best-fitting model). Each model was first simulated using the estimated parameters of each participant (generative parameters) and the same number of trials as there is in the data. Then, the full integration and stimulus-consistency models were fitted to the simulated data and the parameters of each model were extracted (recovered parameters). Supplementary Figure 8 & 9 show the generative parameters plotted against the recovered parameters for the full-integration and stimulus-consistency models, respectively. As can be seen, the quality of the parameter recovery ranges between good to excellent[9].



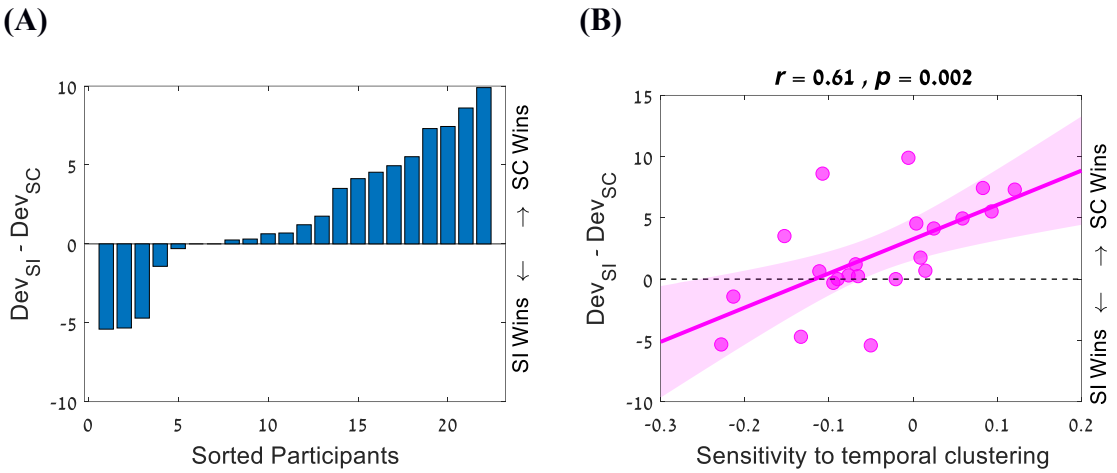

***Supplementary Figure 8.*** *Parameter recovery for the full Integration model. The generative parameters of the full integration model plotted against the recovered parameters for the (A) noise* ($r = .95, p < .001$) *(B) intercept* ($r = .98, p < .001$) *(C) scale* ($r = .98$, $p < .001$) *and (D) asymptote parameters* ($r = .89$, $p < .001$). *Solid lines correspond to linear least-squares fits, and the shaded areas to 95% confidence intervals.*




***Supplementary Figure 9.*** *The generative parameters of the stimulus consistency model plotted against the recovered parameters for the (A) noise* ($r$ = .90, $p$ < .001)*, (B) intercept* ($r = .98, p < .001$)*, (C) scale* ($r = .99, p < .001$)*, (D) asymptote* ($r = .87, p < .001$) *and (E) consistency* ($r = .91, p < .001$) *parameters. Solid lines correspond to linear least-squares fits, and the shaded areas to 95% confidence intervals.*

## Temporal clustering effects

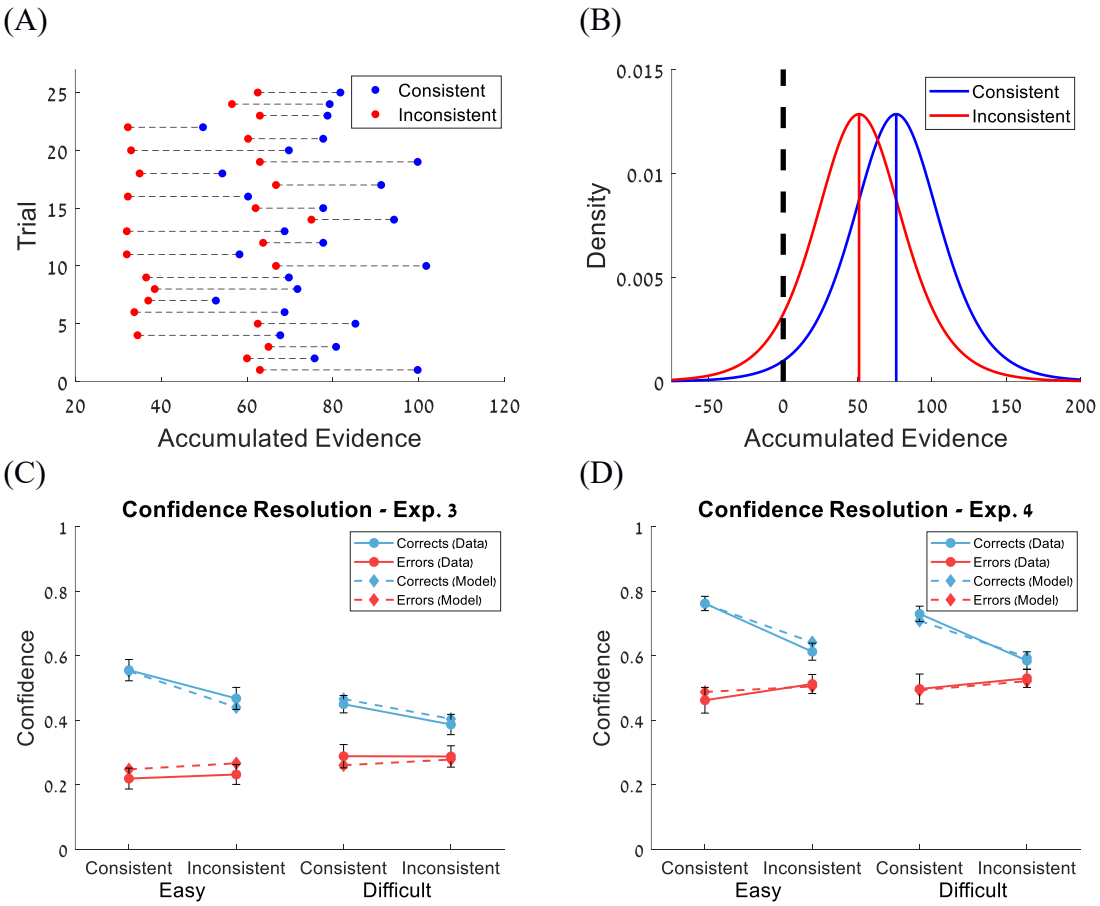
In Experiment 3 we generated consistent and inconsistent trials (Fig. 4A), by manipulating the difference between the number of frames with evidence favoring the two alternatives. This manipulation, however, is correlated with the number of frequent winners, which modulates choices in the selective-integration (SI) model[10]. Thus, although the stimulus-consistency (SC) model provided better fit for the data (Fig. 5E), the consistency effect in Experiment 3 can also be accounted for by the SI model (Fig. 5A). To distinguish between the two models, we used another measure of consistency – LTC (larger temporal cluster of the evidence, see Supplementary Table 1). For each participant, we calculated the LTC measure of each trial, and divided the trials into high/low LTC trials by performing a median split. Then, to quantify the sensitivity of the participants to temporal clustering, we computed the difference in mean accuracy of high and low LTC trials ($accuracy_{high\ LTC} - accuracy_{low\ LTC}$). This difference should be higher for a participant whose bias mechanism is modulated by the consistency of each frame with previous ones, as presenting the positive evidence in an as large cluster as possible enhances the overall integrated evidence. Supplementary Figure 10A shows the differences in goodness of fit (*Deviance* = *-2·LogLikelihood*) between the SI and the SC models (positive values indicate a better fit of the SC model). As shown, the SC model outperformed the SI-model at the group level (see also Fig. 5E) and provided a better fit for the majority of participants. Supplementary Figure 10B shows the correlation between differences in the goodness of fit of the models and the effect of temporal clustering (each point represents a single participant). As shown, the SC model provide a better fit for the data, specifically for participants who showed higher sensitivity to temporal clustering.




***Supplementary Figure 10.*** *(A) Comparison between the SC and SI models across the 22 participants in Experiment 3. (B) Correlation between differences in goodness of fit of the SI and the SC models and the effect of temporal clustering* ($r$ = .61, $p$ = .002)*; each point represents a single participant. Solid lines correspond to linear least-squares fits, and the shaded areas to 95% confidence intervals.*

## Confidence-resolution

In order to account for the differences in confidence resolution in Experiment 3 & 4 (Fig. 5C-D), we simulated the stimulus-consistency model using pairs of consistent and inconsistent trials, that were created using the same generating distributions as in Experiment 3 (that is, with the same evidence content for each pair). Supplementary Figure 11A shows 25 pairs of trials (blue – consistent vs. red – inconsistent), which were simulated using the mean fitted parameters: $\theta_{consistency\ parameter} = 1.75$. The stimulus-consistency model predicts that the bias in favor of consistent evidence would increase the accumulated evidence of consistent compared with inconsistent trials (even though both types of trials have the same evidence). Supplementary Figure 11B shows the distributions of accumulated evidence of consistent and inconsistent trials generated using the stimulus-consistency model (using the mean internal noise obtained in the empirical data, $\beta = .05$). The mean distance from the criterion (0) of correct consistent trials is higher than for correct inconsistent trials, $t(98) = 5.44$, $p < .001$, Cohen's $d = 1.09$, 95% CI 21.80 to 46.88. For inconsistent trials, however, the differences in the mean distances from the criterion of the incorrect consistent and incorrect inconsistent trials did not reach statistical, $t(98) = -0.17$, $p = .87$, Cohen's $d = -0.03$, 95% CI -9.68 to 8.16. Based on the distributions presented in Supplementary Figure 11B we predicted the mean confidence response for each condition of the stimulus consistency (Supplementary Figure 11C-D). To this end, we computed the absolute mean value of the accumulated evidence (that is, distance from 0) of the correct and incorrect responses separately for consistent and inconsistent response. Then, using linear regression we mapped these values to the mean confidence level of each condition (dashed lines).




***Supplementary Figure 11.*** *Confidence-resolution in Experiment 3 & Experiment 4. (A) Twenty-five pairs of trials (consistent and inconsistent) were simulated using the same generating distributions as in Experiment 3 (without internal noise; blue – consistent trials, red - inconsistent). The stimulus-consistency model predicts that the bias in favor of consistent evidence would increase the accumulated evidence of consistent trials compared with inconsistent ones. (B) The distributions of accumulated evidence of consistent (blue) and inconsistent (red) trials simulated using the stimulus-consistency model (with internal noise) in Experiment 3. Solid blue and red vertical bars correspond to the mean of each distribution. (C) Confidence as a function of difficulty and consistency for correct (blue lines) and incorrect (red lines) responses in Experiment 3 (n = 22 participants). Data are shown with solid lines and circle symbols and model predictions are shown with dashed lines and diamond symbols. (D) Same as (C) but for Experiment 4 (n = 25 participants).Data are presented as mean values ± standard error of the mean.*

# Supplementary Methods

## Estimation of non-decision time

The presentation rate in Experiment 1 was 5 Hz (500ms/sample)[1]. Non-decision time (that is, perceptual encoding and response execution) in this experiment was estimated as 250ms. Consequently, if response occurred less than 250ms after the presentation of the last sample, then this sample was excluded from the decision-time calculation. For example, if a participant responded 100ms after the presentation of the eight sample, then the decision-time was defined as seven samples. Additionally, trials in which the response was faster than 250ms, were excluded from further analysis (less than 1% of the data).

The presentation rate in Experiment 2 was 5 Hz (200ms/sample). The higher temporal resolution in this experiment allowed us to estimate non-decision for each participant separately, using the procedure described in ref. [11]. The evidence of the first frame before decision (that is, last frame of each trial) was regressed on choice (left or right alternative) using logistic regression model (hereafter actual-evidence model). The log-likelihood (LL) of the actual-evidence model was compared to the LL of a model in which the evidence was randomly sampled (hereafter random-evidence model). If the LL of the actual-evidence model was lower than the LL of the random-evidence model (indicating a better fit of the former), we assumed that the evidence in the first frame before decision was taken into account whilst deciding. If not, we assumed that it did not affect choice. Evidence was defined as X(n) – Y(n), where X(n) and Y(n) are the last samples of each trial (that is, first samples before decision), drawn from the left and right distributions, respectively. The samples of the random model were drawn from the same generative distributions as in Experiment 2. We used a cut-off of 6 between the deviance values of the models ($\Delta\text{Deviance} = 2\cdot LL_{\text{actual-evidence model}} - 2\cdot LL_{\text{random-evidence model}}$), which is considered strong evidence in favor of the winning model. If the first frame before decision did not affect choice, the second frame before decision was examined, and so on until finding the frame in which the evidence was correlated with choice. Non-decision time was computed as the total number of frames before decision in which the evidence was not associated with choice. The average non-decision time was 194 ± 98ms (0.97 ± 0.49frames), similar to Experiment 1.

## Supplementary References


1. Glickman, M. & Usher, M. Integration to boundary in decisions between numerical sequences. *Cognition* **193**, 104022 (2019).
2. Usher, M. & McClelland, J. L. The time course of perceptual choice: The leaky, competing accumulator model. *Psychol. Rev.* **108**, 550–592 (2001).
3. Hawkins, G. E., Forstmann, B. U., Wagenmakers, E. J., Ratcliff, R. & Brown, S. D. Revisiting the evidence for collapsing boundaries and urgency signals in perceptual decision-making. *J. Neurosci.* **35**, 2476–2484 (2015).
4. Sokolova, M. & Lapalme, G. A systematic analysis of performance measures for classification tasks. *Inf. Process. Manag.* **45**, 427–437 (2009).
5. Boughorbel, S., Jarray, F. & El-Anbari, M. Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. *PLoS One* **12**, e0177678 (2017).
6. Eisen-Enosh, A., Farah, N., Burgansky-Eliash, Z., Polat, U. & Mandel, Y. Evaluation of Critical Flicker-Fusion Frequency Measurement Methods for the Investigation of Visual Temporal Resolution. *Sci. Rep.* **7**, 1 (2017).
7. Ludwig, C. J. H., Gilchrist, I. D., McSorley, E. & Baddeley, R. J. The temporal impulse response underlying saccadic decisions. *J. Neurosci.* **25**, 9907–9912 (2005).
8. Yang, T. & Shadlen, M. N. Probabilistic reasoning by neurons. *Nature* **447**, 1075–1080 (2007).
9. White, C. N., Servant, M. & Logan, G. D. Testing the validity of conflict drift-diffusion models for use in estimating cognitive processes: A parameter-recovery study. *Psychon. Bull. Rev.* **25**, 286–301 (2018).
10. Tsetsos, K. *et al. Economic irrationality is optimal during noisy decision making*. *Proceedings of the National Academy of Sciences of the United States of America* vol. 113 (2016).
11. Malhotra, G., Leslie, D. S., Ludwig, C. J. H. & Bogacz, R. Overcoming indecision by changing the decision boundary. *J. Experiment Psychol. Gen.* **146**, 776–805 (2017).