The suboptimality of perceptual decision making with multiple alternatives

Jiwon Yeon & Dobromir Rahnev

It is becoming widely appreciated that human perceptual decision making is suboptimal but the nature and origins of this suboptimality remain poorly understood. Most past research has employed tasks with two stimulus categories, but such designs cannot fully capture the limitations inherent in naturalistic perceptual decisions where choices are rarely between only two alternatives. We conduct four experiments with tasks involving multiple alternatives and use computational modeling to determine the decision-level representation on which the perceptual decisions are based. The results from all four experiments point to the existence of robust suboptimality such that most of the information in the sensory representation is lost during the transformation to a decision-level representation. These results reveal severe limits in the quality of decision-level representations for multiple alternatives and have strong implications about perceptual decision making in naturalistic settings.
Perception has been conceptualized as a process of inference for over a century and a half. According to this view, the outside world is encoded in a pattern of neural firing and the brain needs to decide what these patterns signify. Hundreds of papers have revealed that this inference process is suboptimal in a number of different ways. However, these papers have almost exclusively employed tasks with only two stimulus categories (though notable exceptions exist). Experimental designs where decisions are always between two alternatives cannot fully capture the processes inherent in naturalistic perceptual decisions where stimuli can belong to many different categories (e.g., which of all possible local species does a particular tree belong to). Therefore, fully understanding the mechanisms and limitations of perceptual decision making requires that we characterize the process of making decisions with multiple alternatives.

One critical difference between perceptual decisions with two versus multiple alternatives is the richness of the sensory information that the decisions are based on. Decisions with two alternatives can be based on the evidence for each of the two categories, or even just the difference between these two pieces of evidence. For example, traditional theories such as signal detection theory and the drift diffusion model postulate that two-choice tasks are performed by first summarizing the evidence down to a single number—the location on the evidence axis in signal detection theory and the identity of the boundary that is crossed in drift diffusion—that is subsequently used for decision making. Thus, the sensory information relevant for the decision in such tasks is relatively simple and could potentially be represented in decision-making circuits without substantial loss of information. However, the relevant sensory information in multi-alternative decisions is more complex because it contains the evidence for each of the multiple alternatives available. Further, this richer sensory information can no longer be summarized in a simple form in decision-making circuits without a substantial loss of information (Fig. 1). However, it is currently unknown whether decision-making circuits can represent the rich information from sensory circuits in the context of multi-alternative decisions or whether the decision-making circuits only represent a crude summary of the sensory representation.

To uncover the decision-level representation of decisions with multiple alternatives, we use discrete stimulus categories in three experiments and stimuli that give rise to a trimodal sensory distribution in a fourth experiment. All experiments feature a condition where subjects pick the dominant stimulus among all of the possible stimulus categories (four different colors in Experiment 1, six different symbols in Experiments 2 and 3, and three different stimulus direction in Experiment 4). Based on these responses, we estimate the parameters of a model describing subjects’ internal distribution of sensory responses (i.e., the activity levels for each stimulus category). We then include conditions where subjects are told to pick between only two alternatives after the offset of the stimulus. These conditions allow us to compare different models of how the sensory representation is transformed into a decision-level representation. To anticipate, we find robust evidence for suboptimality in that decisions in our experiments are based on a summary of the sensory representation thus incurring substantial information loss. These results indicate that perceptual decision-making circuits may not have access to the full sensory representation in the context of multiple alternatives and that significant amount of simplification is likely to occur before sensory information is used for deliberate decisions.

### Results

**Experiment 1.** Subjects saw a briefly presented stimulus consisting of 49 colored circles arranged in a 7 × 7 square (Fig. 2). Each circle was colored in blue, red, green, or white. On each trial, one color was randomly chosen to be dominant (i.e., more frequently presented than the other colors) and 16 circles were painted in that color, whereas the remaining three colors were nondominant and 11 circles were painted in each of those colors. The task was to indicate the dominant color. The experiment featured two different conditions. In the four-alternative condition, subjects picked the dominant color among the four possible colors. In the two-alternative condition, after the offset of the stimulus, subjects were asked to choose between the dominant and one randomly chosen nondominant color. In both conditions, the response screen was displayed with 0-ms delay thus minimizing short-term memory demands. Note that subjects’ task was always the same (to correctly identify the dominant color).

Based on the responses in the four-alternative condition, we estimated the parameters of the sensory distribution representing the activity level for each color. We then considered the predictions for the two-alternative condition of two different models: (1) a population model, according to which perceptual decisions are based on the whole distribution of activities over the four colors, and (2) a summary model, according to which perceptual decisions are based on a summary of the whole distribution. There are a number of ways to create a summary of the whole distribution. However, in the context of this task, the only relevant information is the order of the activation levels from highest to lowest (this order determines how a subject would pick different colors as the dominant color in the two-alternative condition). Other information, such as average activity level, is irrelevant to the task here. Therefore, we first considered an extreme summary model that consists of the activity level for the one color with highest level of activity. Other summary models, in which decision-making circuits have access to the activity levels of the n > 1 colors with highest activity levels, are examined later.

The population and summary models could be easily compared because they make different predictions about performance in the two-alternative task (for a mathematical derivation, see Supplementary Methods). Indeed, the models make the same prediction when the dominant color gives rise to the highest activity level (Fig. 3a) and when the alternative option given to the subject happens to have the highest activity (Fig. 3b), but diverge when the highest activity is associated with a color that is not among the two options with the population model predicting a higher performance level (Fig. 3c).

The difference between the two models could be seen in the actual model predictions. Indeed, based on the performance in the four-alternative condition (average accuracy = 69.2%, chance level = 25%), the population and summary models predicted an average accuracy of 84.2% and 79.7% in the two-alternative condition, respectively. Compared to the actual subject performance (average accuracy = 78%), the population model overestimated the accuracy in the two-alternative conditions for 29 of the 32 subjects (average difference = 6.21%; t(31) = 8.19, p = 3.02 × 10⁻⁹, 95% CI = [4.7%, 7.8%]). Surprisingly, the summary model also overestimated the accuracy in the two-alternative condition but the misprediction was much smaller (average difference = 1.72%; t(31) = 2.35, p = 0.025, 95% CI = [0.2%, 3.2%]) (Fig. 4a). Indeed, the absolute error of the predictions of the population model was significantly larger than for the summary model (average difference = 2.93%; t(31) = 5.65, p = 3.34 × 10⁻⁶, 95% CI = [1.9%, 4%]). Overall, the summary model predicted the accuracy in the two-alternative condition better than the population model for 26 of the 32 subjects (Fig. 4b).
We further compared the models’ fits to the whole distribution of responses. We found that the Akaike Information Criterion (AIC) favored the summary model by an average of 24.30 points (Fig. 4c), which corresponds to the summary model being $1.89 \times 10^5$ times more likely for the average subject. Across the whole group of 32 subjects, the total AIC difference was thus 777.63 points, corresponding to the summary model being $7.26 \times 10^{168}$ times more likely in the group. Note that since the population and summary models had the same number of parameters, the same results would be obtained regardless of the exact metric employed (e.g., the BIC differences would be exactly the same; see “Methods” for details).

Finally, we constructed and tested four additional models. The first two models postulated that decision-making...
circuits have access to the two or three highest activations of the sensory distribution (two-highest and three-highest models, respectively). These models could thus be seen as intermediate options between the summary and population models. Two other models postulated that subjects choose either two or three stimulus categories to attend to and then make their decisions based on a full probability distribution over the activity levels of the attended categories (two-attention and three-attention models; see Supplementary Methods for details). We found that all of these models were outperformed by the summary model (Supplementary Notes and Supplementary Figs. 1, 6a–d).

Fig. 3 Predictions of the two models for choices in the two-alternative condition. The population model (left panels) assumes that decision-making circuits have access to the activity levels associated with each of the four colors (four gray bars), whereas the summary model (right panels) assumes that decision-making circuits only have access to the highest activity level (single gray bar). In all examples, the dominant circle is white, and subjects are given a choice between white and green. a When the highest activity happens to be at the dominant color, both models predict that the subject would correctly choose the dominant color. b When the highest activity happens to be at the alternative color, both models predict that the subject would incorrectly choose the alternative color. c The predictions of the two models diverge when the highest activity is associated with a color other than the two presented alternatives. In such cases, the activation for the dominant color is likely to be higher than for the alternative color, so according to the population model, subjects would ignore the color with the highest activity (blue color in the example here) and correctly pick the dominant color in the majority of the trials. However, according to the summary model, subjects have no information about the activation levels for the dominant and the alternative colors and would thus correctly pick the dominant color on only 50% of such trials.
Population model (predicted accuracy differed by only 7.9 points. On average, the summary model had the population model was favored over the summary model 25 points (Fig. 5d). The AIC values of the one subject for whom the difference in AIC values in all these nine subjects was larger than the summary model was preferred by nine of our ten subjects and the summary model provided better prediction of the accuracy in the data. Each dot represents one subject. The gray horizontal lines at ±3 and ±10 indicate common thresholds for suggestive and strong evidence for one model over another. The red triangle indicates the average AIC difference. The summary model provided a better fit than the population model for 30 of the 32 subjects.

Experiment 2. The results from Experiment 1 strongly suggest that within the context of our experiment, decision-making circuits do not represent the whole sensory distribution but only a summary of it. We sought to confirm and generalize these findings in two additional, preregistered experiments ([osf.io/dr89k]). For Experiment 2, we made several modifications: (1) we changed the colored circles to symbols, (2) we raised the number of stimulus categories from four to six, and (3) we significantly increased the number of trials per subject in order to obtain stronger results at the individual-subject level. Specifically, we presented the six symbols (?, #, $, %, +, and >) such that the dominant symbol was presented 14 times and each nondominant symbol was presented 7 times (Fig. 5a). The 49 total symbols were again arranged in a 7 × 7 grid. Each subject completed a total of 3000 trials that included equal number of trials of a six-alternative condition and a two-alternative condition that were equivalent to the four- and two-alternative conditions in Experiment 1.

Just as in Experiment 1, we computed the parameters of the sensory representation using the trials from the six-alternative condition (average accuracy = 50.5%, chance level = 16.7%) and used these parameters to compare the predictions of the population and summary models for the two-alternative condition. We found that the average accuracy in the two-alternative condition (71.6%) was slightly underestimated by the summary model (predicted accuracy = 70.1%; t(9) = −2.76, p = 0.022, 95% CI: [−2.7%, −0.3%]) but was again significantly overestimated by the population model (predicted accuracy = 77.5%; t(9) = 9.41, p = 5.92 × 10−6, 95% CI [4.5%, 7.3%]) (Fig. 5b). Individually, the summary model provided better prediction of the accuracy in the two-alternative condition for nine out of the ten subjects (Fig. 5c).

Further, we compared the population and summary models’ fits to the whole distribution of responses. We found that the summary model was preferred by nine of our ten subjects and the difference in AIC values in all these nine subjects was larger than 25 points (Fig. 5d). The AIC values of the one subject for whom the population model was favored over the summary model differed by only 7.9 points. On average, the summary model had an AIC value that was 57.79 points lower than the population model corresponding to the summary model being 3.55 × 10^{12} times more likely for the average subject. Across the whole group of ten subjects, the total AIC difference was thus 577.94 points, corresponding to the summary model being 3.14 × 10^{125} times more likely in the group. Finally, we found that the additional four models were again outperformed by the summary model (Supplementary Notes and Supplementary Figs. 2, 6e–h).

Experiment 3. Taken together, Experiments 1 and 2 suggest that in the context of multi-alternative decisions, the system for deliberate decision making may not have access to the whole sensory representation. This conclusion is based on experiments that differed in the nature of the stimulus, the number of stimulus categories, and the amount of trials that subjects performed. Nevertheless, both Experiments 1 and 2 relied on the same design of comparing four- (or six-) and two-alternative conditions. Therefore, to further establish the generality of our results, in Experiment 3 we employed a different experimental design. We used the same stimulus as in Experiment 2 and presented all six alternatives on every trial, but additionally gave subjects the opportunity to provide a second answer on about 40% of error trials (Fig. 6a). Using the performance on the first answer, we compared the predictions of the population and summary models for the second answers.

The population model makes a clear prediction about the second answer—subjects should choose the stimulus category with the highest activation from among the remaining five options. The second answer will thus have relatively high accuracy because the presented stimulus category is likely to produce one of the highest activity levels (Supplementary Fig. 3a). On the other hand, the summary model only features information about the stimulus category with the highest activity. Once that stimulus category is chosen as the first answer, the model postulates that the subject does not have access to the activations associated with the other stimulus categories. Given this representation, subjects could adopt at least two different response strategies. One possible strategy is for the subject to make their second answer at random, which would result in chance level (20%) performance. We call this the Summary & Random Choice model (Supplementary Fig. 3b). However,
another possibility is for the subject to make the second answer strategically. One available strategy is for the subject to pick the stimulus category of a randomly recalled symbol from the $7 \times 7$ grid. Given that subjects inspected the stimuli for 500 ms, they could easily remember one location with a symbol other than the one they picked for their first answer. We call this the Summary & Strategic Choice model (Supplementary Fig. 3c). According to this model, the second answer will be correct on $\frac{14}{42} \approx 0.333$ or 33.3% of the time. Conversely, each of the four remaining incorrect categories will be chosen on $\frac{7}{42} \approx 0.167$ or 16.7% of the time (Supplementary Fig. 4).

To adjudicate between these three models, we first examined subjects’ accuracy on the first answer. Subjects responded correctly in their first answer on 50.7% of the trials (chance level = 16.7%). Using this performance, we computed the parameters of the sensory representation as in Experiments 1 and 2 in order to generate the model predictions for the second answer. We found that task accuracy for the second answers was 29.6%. This value was greatly overestimated by the population model, which predicted accuracy of 40.9% ($t(9) = 7.04$, $p = 6.09 \times 10^{-5}$, 95% CI = [7.7%, 15%]) (Fig. 6b). On the other hand, the Summary & Random Choice model greatly underestimated the observed accuracy (predicted accuracy = 20%; $t(9) = -5.55$, $p = 3.55 \times 10^{-4}$, 95% CI = [−13.5%, −5.7%]). Finally, the Summary & Strategic Choice model produced the most accurate prediction (predicted accuracy = 33.3%; $t(9) = 2.18$, $p = 0.057$, 95% CI = [−0.1%, 7.7%]). On an individual subject level, the population model overestimated the accuracy of the second answer for all ten subjects, the Summary & Random Choice model underestimated the accuracy of the second answer for all of the ten subjects, whereas the Summary & Strategic Choice model was best calibrated overestimating the accuracy of the second answer for seven subjects and underestimating it for the remaining three subjects (Fig. 6c).

Formal comparisons of the model fits to the full distribution of responses for the second answers demonstrated that the population model provided the worst overall fit (Fig. 6d). Indeed, the population model resulted in AIC values that were higher than the Summary & Random Choice model by an average of 18.05 points (corresponding to $8.29 \times 10^{2}$-fold difference in likelihood in the average subject) and a total of 180.46 points (corresponding to $1.53 \times 10^{9}$-fold difference in
The population model underperformed the Summary & Strategic Choice model even more severely (average AIC difference = 37.29 points, corresponding to $1.25 \times 10^8$-fold difference in likelihood in the average subject; total AIC difference = 372.93 points, corresponding to $9.57 \times 10^{80}$-fold difference in likelihood in the group). Lastly, the Summary & Strategic Choice model also outperformed all of the additional models (Supplementary Results and Supplementary Fig. 6i–l). Thus, just as Experiments 1 and 2, Experiment 3 provides strong evidence that in the context of multi-alternative decisions, decision-making circuits only contain a summary of the sensory representation.

**Experiment 4.** To adjudicate between the population and summary models, Experiments 1–3 employed discrete stimulus categories (Fig. 1a). However, it remains possible that the results of these experiments would not generalize to stimuli represented on a continuous scale. To address this issue, we performed a fourth experiment that employed a feature (dot motion) represented on a continuous scale (degree orientation). We adapted the design of Treue et al. in which groups of dots slid transparently across one another. Specifically, we presented moving dot stimuli where three sets of dots moved in three different directions. As Treue et al. noted, these stimuli produce the subjective experience of seeing three distinct surfaces sliding across each other and

---

**Fig. 6 Task and results for Experiment 3.** a The same stimuli as in Experiment 2 were used in Experiment 3 but the task was slightly different. Subjects always reported the dominant symbol among all six alternatives. However, on 40% of the trials in which they gave a wrong answer, subjects were given the opportunity to make a second guess. b Mean accuracy for the second answer observed in the actual data (white bar), predicted by the population model (light gray bar), predicted by the Summary & Random Choice model (dark gray bar), and predicted by the Summary & Strategic Choice model (black bar). The predictions of the three models were derived based on subjects’ first answers. All $p$ values are derived from two-sided paired t tests. Error bars represent SEM, $n = 10$. c Individual subjects’ differences in the accuracy of the second answer between each model’s prediction and the observed data. d AIC difference between the population and the two summary models. Positive AIC values indicate that the summary model provides a better fit to the data. Each dot represents one subject. The red triangle indicates the average AIC difference.
the summary model predicted the observed accuracy well for nine out of 11 subjects (Fig. 7c). Two models, the summary model predicted the observed task accuracy better for nine out of 11 subjects (Fig. 7c). The total AIC difference across all subjects was 60.15 points lower for the summary model, corresponding to the summary model being $1.15 \times 10^{13}$ times more likely than the population model.

Discussion

We investigated whether the decision-level representation in decisions with multiple alternatives consists of a copy of the sensory representation or only a summary of it. We performed four experiments with either discrete stimulus categories or continuous stimuli producing multimodal distributions. The results across all experiments showed that the population model that assumes no loss of information from sensory to decision-making circuits did not provide a good fit to the data. Instead, the summary model, which assumes that decision-making circuits represent a reduced form of the sensory distribution, consistently provided a substantially better fit. These results strongly suggest that deliberate decision making for multiple alternatives only has access to a summary form of the sensory representation.

To the best of our knowledge, the current experiments are the first to address the question of whether complex sensory codes for a single stimulus can be accurately represented in decision-making circuits. Prior studies have convincingly demonstrated that humans can form complex, non-Gaussian and even bimodal priors over repeated exposures to a given stimulus. However, it should be emphasized that this previous research has focused on the ability to learn a prior over many trials and did not

![Fig. 7 Task and results for Experiment 4 which uses stimuli on a continuous scale.](image)

(a) Three sets of dots moved in three different directions separated by 120°. Similar to Experiments 1–3, one of the three sets of dots had more dots (i.e., dominant direction) compared to the other two sets (i.e., nondominant directions). Each trial began with a fixation cross followed by the moving dot stimulus presented for 500 ms. The response screen was presented immediately after the offset of the moving dots and randomly assigned a stimulus-response mapping on each trial. Similar to Experiments 1 and 2, subjects picked the dominant direction of motion between all three directions (three-alternative condition) or between the dominant direction and one randomly chosen nondominant direction (two-alternative condition). (b) Mean accuracy in the two-alternative condition observed in the actual data (white bar) and predicted by the population (light gray bar) and summary (dark gray bar) models. The predictions for both models were derived based on the data in the three-alternative condition. All $p$ values are derived from two-sided paired $t$ tests. Error bars represent SEM, $n = 11$. Individual subjects’ differences in the accuracy of the two-alternative condition between the predictions of each of the two models and the observed data. (c) AIC difference between the population and summary models. Positive AIC values indicate that the summary model predicts the observed data better. The red triangle indicates the average AIC difference. The summary model fits better than the population model for nine out of the 11 subjects.
examine the ability to use the sensory representation produced by a single stimulus on a single trial.

Why is it that decision-making circuits do not maintain a copy of the full sensory representation? While our study does not directly address the origins of this suboptimality, it is likely that the reason for the suboptimality lies in decision-making circuits having markedly smaller capacity than sensory circuits. One potential reason for the smaller capacity is that decision-making circuits have to be able to represent many different features (e.g., orientation, color, shape, object identity, etc.) each of which is processed in dedicated sensory regions. According to this line of reasoning, the generality inherent in decision-level representations necessitates that detail present in sensory cortex is lost. A related reason for the information loss in decision-making circuits is that the representations in such circuits often need to be maintained over a few seconds and therefore are subject to the well-known short-term memory decay.

For example, even though all conditions in our task were shown with a 0-ms delay, information may need to be passed serially from sensory to decision-making circuits thus inherently inducing short-term memory demands. According to this line of reasoning, even if decision-making circuits can represent a full copy of the information in sensory cortex, that copy will decay as soon as it begins to be assembled and thus information loss with necessarily accrue before the representation can be used for subsequent computations. Therefore, both the necessary generality of decision-level representations and the inherent limitations of short-term memory likely contribute to the sparse representations in decision-making circuits.

If decision-making circuits indeed only have access to a summary form of the sensory representation, does that mean that absolutely no computations can be based on the complete sensory representations? There is evidence that decisions can take into account the full sensory representation both in simple situations where only two alternatives are present and in cases of automatic multisensory integration. In general, automated computations performed directly on the sensory representations may use the whole sensory representation. Therefore, certain types of decision making can be performed via mechanisms that do indeed take advantage of the entire sensory representation with either no or minimal loss of information. However, it is likely that such computations are restricted to either very simple decisions or processes that are already automated. On the other hand, the last stage of decision-making supporting nonautomatic, flexible, and deliberate decisions (which require short-term memory maintenance) only has access to a summary of the sensory representation.

Our findings can be misinterpreted as suggesting that complex visual displays are represented as a point estimate: that is, the decision-level representation features only the best guess of the system (e.g., 60° orientation, red color, or the plus symbol). The possibility of a decision-level representation consisting of a single point estimate has been thoroughly debunked. For example, a point estimate does not allow us to rate how certain we are in our decision because we lack a sense of how uncertain our point estimate is. Given that humans and animals can use confidence ratings to judge the likely accuracy of their decisions, decision-making circuits must have access to more than a point estimate of the stimulus.

It should therefore be clarified that our summary model does not imply that decision making operates on point estimates. Indeed, as conceptualized in Fig. 1, the summary model assumes that subjects have access to both the identity of the most likely stimulus category (e.g., the color white) and the level of activity associated with that stimulus category. The level of activity can then be used as a measure of uncertainty, and confidence levels can be based on this level. Such confidence ratings will be less informative than the perceptual decision, which is exactly what has been observed in a number of studies. In addition, this type of confidence generation may explain findings that confidence tends to be biased towards the level of the evidence for the chosen stimulus category and tends to ignore the level of evidence against the chosen category. Thus, a summary model, consisting of the identity of the most likely stimulus and the level of activity associated with this stimulus, appears to be broadly consistent with findings related to how people compute uncertainty and is qualitatively different than a decision-level representation consisting only of a point estimate.

Another important question concerns whether any additional information is extracted from the sensory representation beyond what is assumed by the summary model. It is well known that humans can quickly and accurately extract the gist of a scene, as well as the statistical structure of an image. Therefore, it appears that rich information is extracted during the time when the stimulus is being viewed. In fact, this information often goes beyond the extraction of just the identity of the most likely stimulus and the level of activity associated with this stimulus assumed by our summary model. For example, our moving dots stimulus in Experiment 4 resulted in the perception of three surfaces sliding on top of each other. This means that what was extracted in that experiment was the approximate location of each of the three peaks of the trimodal sensory distribution. Similarly, the subjects in our Experiment 1 were certainly aware that four different colors were presented in each display and would have noticed if we ever presented additional colors. Subjects, therefore, had access to the identity of the different colors presented even though they did not have information about the activity level associated with each color. Thus, our summary model is likely to be an oversimplification of the actual representation used for decision-making. This point is further underscored by the fact that when predicting the accuracy in the two-alternative condition, the summary model showed a slightly lower correlation with the true accuracy than our model. In other words, even though CNNs extract complex representations in their early layers, the networks do not perform decision making based directly on these sensory-like representations in their early layers.

In conclusion, we found evidence from one exploratory (Experiment 1) and two preregistered (Experiments 2 and 3) studies that deliberate decision making for discrete stimulus categories is performed based on a summary of, rather than the whole, sensory representation. A final study (Experiment 4) extended these results to stimuli that give rise to continuous multimodal distributions. Our findings demonstrate that flexible computations may not be performed using the sensory activity itself but only a summary form of that activity.

Methods

Subjects

A total of 63 subjects participated in the four experiments. Experiment 1 had 22 subjects (15 females, mean age = 20.13, SD = 2.21, range = 18–29), Experiment 2 had 10 subjects (7 females, mean age = 20.5, SD = 3.06, range = 18–28), Experiment 3 had 10 subjects (5 females, mean age = 20.8, SD = 3.55,
range = 18–28), and Experiment 4 had 11 subjects (6 females, mean age = 21.45, SD = 2.5, range = 18–28). Each subject participated in only one experiment. All subjects were given a total of 3000 trials each. The study was approved by the Georgia Tech Institutional Review Board (H15308).

Apparatus and experiment environment. The experiments stimuli were presented on a 21.5-inch iMac monitor in a dark room. The distance between the monitor and the subjects was 60 cm. The stimuli were created in MATLAB, using Psychtoolbox 3.4.2.

Experiment 1. The stimulus consisted of 49 circles colored in four different colors — red, blue, green, and white — presented in a 7 × 7 grid on black background. The diameter of each colored circle was 0.24° and the distance between the centers of two adjacent circles was 0.6°. The grid was located at the center of the screen. On each trial, one of the four colors was dominant — it was featured in 16 different locations— whereas the other three colors were nondominant and were featured in 11 locations each. The exact locations of each color were pseudo-randomly chosen so that each color was presented the desired number of times. A trial began with a 500-ms fixation followed by 500-ms stimulus presentation. Subjects then indicated the dominant color in the display and provided a confidence rating without time pressure.

There were three different conditions in the experiment. In the first condition, subjects could choose any of the four colors (four-alternative condition). In the second condition, after the stimulus offset subjects were asked to choose between only two options that were not announced in advance — one was always the correct dominant color and the other was a randomly selected nondominant color (two-alternative condition). Finally, in the third condition, subjects were told in advance which two colors will be queried at the end of the trial (advance warning condition). For the purposes of the current analyses, we only analyzed the four- and two-alternative conditions. The advanced warning condition and the confidence ratings were not analyzed.

Subjects completed six runs, each consisting of three 35-trial blocks (for a total of 630 trials). The three conditions used in the experiment were blocked such that one block in each run consisted entirely of trials from one condition and each run included one block from each condition. Subjects were given 15-5 breaks between blocks and untimed breaks between runs. Before the start of the main experiment, subjects completed a training session where they completed 15 trials per condition with trial-to-trial feedback, and another 15 trials per condition without trial-to-trial feedback. No explicit feedback was provided during the main experiment in any of Experiments 1–4 though the presence of second answers in Experiment 3 served as a form of feedback that those specific trials were wrong. We did not hypothesize that the presence or absence of feedback would alter the results in a systematic way and therefore chose to withhold feedback as in our previous experiments.43,44

Experiment 2. Following our exploratory analyses on the data from Experiment 1, we preregistered two additional experiments (Experiments 2 and 3) [osf.io/dr9bk/]. These experiments were designed to generalize the results from Experiment 1 and to obtain stronger evidence for our model comparison results on the individual subject level. Consequently, we had fewer number of subjects in Experiments 2 and 3 but each subject completed many more trials. We ended up making three deviations from the preregistration. (1) we included a different number of locations for nondominant items (the preregistration wrongly indicated a number that is impossible given the number of categories and total number of characters), (2) we used a dot for the fixation even though the preregistration indicated that we would use a crosshairline, and (3) we tested additional models (the preregistration only included the population and summary models from Experiment 1).

The stimulus in Experiments 2 and 3 consisted of 49 characters from among six possible symbols (i.e., _, $, & , %, +, >) presented in a 7 × 7 grid. The symbols were chosen to be maximally different from each other. The symbols’ width was 0.362° on average and height was 0.66° on average. The distance between two centers was 1.1°. The symbols were presented in the background. On each trial, one of the six symbols was dominant — it was featured in 14 different locations— whereas the other five were nondominant and were featured in 7 locations each. The exact locations in the 7 × 7 grid where each symbol was displayed were pseudo-randomly chosen so that each symbol was presented the desired number of times. Each trial began with a 500-ms fixation, followed by a 500-ms stimulus presentation. The stimuli were then masked for 100 ms with a 7 × 7 grid of ellipsoid-shaped images consisting of uniformly distributed noise pixels. Each ellipsoid had a width of 0.54° and height of 0.95°, ensuring that it entirely covered the symbol in the display without time pressure.

Each subject completed three sessions of the experiment on different days. Each session started with a short training session. On the first day, subjects completed six blocks of 40 trials with the three-alternative condition. The proportion of dots moving in the dominant direction was initially set to 60%. After each block, the proportion of dots moving in the dominant direction was adjusted by half of a previous proportion change. After the sixth block, the proportion of dots moving in the dominant direction was adjusted by half of a previous proportion change. After the sixth block, the proportion of dots moving in the dominant direction was adjusted by half of a previous proportion change. After the sixth block, the proportion of dots moving in the dominant direction was adjusted by half of a previous proportion change. After the sixth block, the proportion of dots moving in the dominant direction was adjusted by half of a previous proportion change. After the sixth block, the proportion of dots moving in the dominant direction was adjusted by half of a previous proportion change.

Model development for Experiments 1–3. We developed and compared two main models of the decision-level representation. According to the population model, decision-making circuits have access to the whole sensory representation. On the other hand, according to the summary model, decision-making circuits only have the access to a summary of the sensory representation but not to the whole sensory representation.

In order to compare the population and summary models, we first had to develop a model of the sensory representation. We created this model using the four- and six-alternative conditions in Experiments 1 and 2, and the first answer in Experiment 3. The population and summary models were then used to make predictions about the two-alternative condition in Experiments 1 and 2, and the second answer in Experiment 3. These predictions were made without the use of any extra parameters.

We created a model of the sensory representation for Experiment 1 as follows. We assumed that each of the four types of stimuli (red, blue, green, or white being the dominant color) produced variable across-trial activity corresponding to each of the four colors. We modeled this activity as Gaussian distributions whose mean (μ) is a free parameter and variance is set to one. However, in our experiments, the
perceptual decisions only depended on the relative values of the activity levels and not on their absolute values. In other words, adding a constant to all four \( \mu \)s for a given color would result in equivalent decisions. Therefore, without loss of generality, we set the mean for the activity corresponding to each dominant color as 0. This procedure resulted in 12 different free parameters such that for each of the four possible dominant colors there were 3 \( \mu \)s corresponding to each of the nondominant colors. Finally, we included an additional parameter that models the lapse rate. Note that the inclusion of a lapse rate parameter has a greater influence on percent correct in the two-alternative compared to the four-alternative condition because overall performance is higher in the two-alternative condition. Therefore, introducing a lapse rate parameter favors the population model by leading to predictions of lower performance in the two-alternative condition (which helps the population model since it consistently predicts higher performance than what was empirically observed).

The sensory representation was modeled in a similar fashion in Experiments 2 and 3. In both cases, the model was created based on subjects choosing between all possible stimulus pairs. The model of the sensory representation in Experiments 2 and 3 thus had 30 free parameters related to the sensory activations (for each of the six possible dominant symbols there were five \( \mu \)s corresponding to each of the nondominant symbols) and an additional free parameter for the lapse rate.

We modeled the actions introduced by each stimulus type separately to capture potential relationships between different colors or symbols (e.g., some color pairs may be perceptually more similar than others). However, we re-did all analyses using the simplifying assumption that when a color is nondominant, that color has the same \( \mu \) regardless of which the dominant color is. This assumption allowed us to significantly reduce the number of parameters in our model of the sensory representation. In this alternative model of the sensory representation, the mean activity for each color/symbol was determined only based on whether that color/symbol was dominant or not. Therefore, we included two free parameters for each color/symbol. However, because of the issue described above (adding a constant to all \( \mu \)s in a given experiment would result in identical decisions), we fixed one of the \( \mu \)s to 0. This modeling approach reduced the total number of free parameters to eight in Experiment 1 (seven \( \mu \)s and a lapse rate) and 12 free parameters in Experiments 2 and 3 (11 \( \mu \)s and a lapse rate). This modeling approach produced virtually the same results (Supplementary Fig. 5).

Lastly, we considered two different instantiations of the summary model for Experiment 3. In the first instantiation, which we refer to as the Summary & Random Choice model, it is assumed that when the first answer is wrong, then the subject would randomly pick a second answer among the remaining options. In the second instantiation, which we refer to as the Summary & Strategic Choice model, it is assumed that when the first answer is wrong, then the subject would pick the stimulus category of a randomly recalled symbol from the original 7 \times 7 grid that is different from the stimulus category chosen with the first answer. According to this model, the subject would pick the second answer correctly 33.3% of the time and each incorrect symbol will be chosen 16.7% of the time (Supplementary Fig. 4).

**Model development for Experiment 4.** Experiment 4 employed moving dots and required subjects to indicate the dominant direction of motion. Therefore, unlike the tasks in Experiments 1–3 that were based on discrete categories of stimuli, the task in Experiment 4 featured a continuous variable (direction of motion, varying from 0° to 360°). However, despite the continuous nature of the stimulus, the three directions of motion could be easily identified implying that the stimulus resulted in a tridimensional sensory distribution6. This allowed us to use the same modeling approach from Experiments 1–3 by essentially treating the three motion directions as discrete stimuli. We again developed a model of the sensory representation that was fit to the three-alternative condition. Unlike Experiments 1–3 where the stimulus pairs were fixed, here the motion direction of the dominant stimulus was chosen randomly (from 0° to 360°) on every trial. Therefore, the model only had parameters for the heights of the nondominant and dominant directions of motion. Because, just as in the previous experiments, adding a constant to all parameters would result in identical decisions, the parameters for the nondominant direction was fixed thus leaving us with a single free parameter. Once the model was fit to the data from the three-alternative condition, the population and summary models had no free parameters when applied to the data from the two-alternative condition.

**Model fitting and model comparison.** For all four experiments, we fit the models to the data as previously15 using a maximum likelihood estimation approach. The models were fit to the full distribution of probabilities of each response type contingent on each stimulus type:

\[
\log \text{likelihood} = \sum_{ij} \log p_{ij} \times n_{ij},
\]

where \( p_{ij} \) is the predicted probability of giving a response \( i \) when stimulus \( j \) is presented, whereas \( n_{ij} \) is the observed number of trials where a response \( i \) was given when stimulus \( j \) was presented. We give formulas for computing \( p_{ij} \), in a simplified model without a lapse rate in the Supplementary Methods. Because the analytical expressions to obtain \( p_{ij} \) are difficult to compute, we derived the model behavior for every set of parameters by numerically simulating 100,000 individual trials with that parameter set. Model fitting was done by finding the maximum-likelihood parameter values using simulated annealing19. Fitting was conducted separately for each subject.

Based on the parameters of the model describing the sensory representation, we generated predictions for the two-alternative condition (Experiments 1, 2, and 4) and the second answer (Experiment 3) for both the population and summary models. These predictions contained no free parameters. To compare the models, we computed AIC based on the log-likelihood (\( \log(L) \)) for each model using the formula:

\[
AIC = -2 \times \log(L) + 2 \times k,
\]

where \( k \) is the number of parameters of a model. Because both the population and summary models had no free parameters, this formula simplifies to \( AIC = -2 \times \log(L) \). Note other measures, such as the AIC corrected for small sample sizes (AICc) given by the formula:

\[
AICc = -2 \times \log(L) + 2 \times k \times \frac{k + 1}{n - k - 1}
\]

where \( n \) is the total number of observations, and the Bayesian Information Criterion (BIC) given by the formula:

\[
BIC = -2 \times \log(L) + k \times \log(n)
\]

would result in the exact same values as AIC for \( k = 0 \). We chose to report AIC values instead of the raw \( \log(L) \) values because of their wider usage and larger familiarity but all conclusions would remain the same if the raw \( \log(L) \) values are considered. Note that lower AIC values correspond to better model fits.

**Statistical tests.** All statistical tests reported are two-sided paired \( t \) tests.

**Data availability**

Model fitting data and results of the four experiments are available at the OSF repository (https://osf.io/d2bb9/)50. In addition, the data from Experiment 1 were made available on the Confidence Database (https://osf.io/s46pr/)13. A reporting summary for this article is available as a Supplementary Information file. Source data are provided with this paper.

**Code availability**

The analysis codes for all four experiments are available at the OSF repository (https://osf.io/d2bb9/)50.
