Supporting Information

Giguère and Love 10.1073/pnas.1219674110

SI Results

Empirical Data. For all experiments, training trials with extreme response times (under 150 ms or over 10 s) were eliminated from the training analyses (experiment 1, 1.2% of trials; experiment 2, 0.6%; experiment 3, 0.4%). Excluding outliers did not change the pattern of analyses. No trials were eliminated from the test analyses.

Experiments 1 and 2. In experiment 1, participants were either trained on a highly probabilistic category structure (Fig. 3*A*) or an idealized structure with reduced variance (Fig. 3*B*). In experiment 2, participants either received feedback based on the same highly probabilistic structure used in experiment 1 or a purely deterministic structure (i.e., the idealized condition; Fig. 3*C*).

Training phase. In experiment 1, the proportion of trials in agreement with the optimal classifier was significantly higher for participants in the idealized condition (mean = 78.64%, SE = 2.05%) than for those in the actual condition [mean = 63.62%, SE = 1.59%; independent samples two-tailed t test, $t(84) = 5.814$, $P < 0.001$]. Median response times for trials in accord with the optimal classifier did not differ between the actual (mean = 676 ms, SE = 37 ms) and idealized (mean = 635 ms, SE = 22 ms) conditions [independent samples two-tailed t test, $t(84) = 0.941$, $P = 0.350$].

In experiment 2, the proportion of trials in agreement with the optimal classifier was significantly higher for participants in the idealized condition (mean = 84.85%, SE = 0.83%) than for those in the actual condition [mean = 64.94%, SE = 1.8%; independent samples two-tailed t test, $t(91) = 10.384$, $P < 0.001$]. Median response times for trials in accord with the optimal classifier did not significantly differ between conditions [actual condition: mean = 731 ms, SE = 52 ms; idealized condition: mean = 624 ms, SE = 23 ms; independent samples two-tailed t test, $t(91) = 1.957$, $P = 0.053$].

Test phase. Figs. 4 and 5 show the test phase results. In experiment 1, participants in the idealized condition (mean = 84.21%, SE = 3.31%) produced significantly more responses in accord with the optimal classifier than those from the actual condition [mean = 70.87%, SE = 3.43%; independent samples two-tailed t test, $t(84) = 2.794$, $P = 0.006$]. Median response times for trials in accord with the optimal classifier were significantly slower in the actual (mean = 657 ms, SE = 54 ms) than in the idealized condition [mean = 501 ms, SE = 13 ms; independent samples two-tailed t test, $t(84) = 2.759$, $P = 0.007$].

In experiment 2, the proportion of trials in accord with the optimal classifier was significantly higher in the idealized condition (mean = 92.01%, SE = 0.63%) than in the actual condition [mean = 74.36%, SE = 2.86%; independent samples two-tailed t test, $t(91) = 6.333$, $P < 0.001$]. Median response times for trials in accord with the optimal classifier were significantly slower in the actual (mean = 642 ms, SE = 39 ms) than in the idealized condition [mean = 507 ms, SE = 16 ms; independent samples two-tailed t test, $t(91) = 3.321$, $P = 0.001$].

When studying individual test response patterns, inconsistencies were defined as cases in which neighboring stimuli were classified in opposite categories. Hence, an optimal classifier would produce a single inconsistency. In experiment 1, participants in the idealized condition (mean = 8.31, SE = 1.19) produced significantly fewer inconsistencies than those from the actual condition [mean = 15.41, SE = 1.08; independent samples two-tailed t test, $t(84) = 4.433$, $P < 0.001$]. In experiment 2, participants in the idealized condition (mean = 6.49, SE = 0.49) also produced significantly fewer inconsistencies than those from the actual condition [mean = 15.41, SE = 1.39; independent samples two-tailed t test, $t(91) = 6.288$, $P < 0.001$].

Experiment 3. In experiment 3, participants made decisions in a complex real-world domain, namely that of professional sports forecasting. Participants predicted the outcome of Major League baseball games following training on either actual game outcomes or idealized outcomes based on the rank of each team in the training sample. Baseball results constitute a fairly noisy domain, as is shown in Table S1.

Training phase. To compare training performance across conditions, data were analyzed in light of the proportion of responses in agreement with rank-order feedback. By this metric, participants in the idealized condition (mean = 83.39%, SE = 1.42%) significantly outperformed those from the actual outcome condition [mean = 56.12%, SE = 0.95%; dependent samples two-tailed t test, $t(41) = 18.165$, $P < 0.001$]. Median response times for trials in accord with rank-order feedback did not differ between the idealized (mean = 1,397 ms, SE = 48 ms) and actual (mean = 1,354 ms, SE = 74 ms) conditions [dependent samples two-tailed t test, $t(41) = 0.538$, $P = 0.594$].

Test phase. Fig. 5 shows the test-phase results. At test, the score awarded for a specific trial was equal to the proportion of non-training games won by the chosen team (using actual results from the database). The rationale behind this scoring method is that it is equivalent to testing participants on all remaining season games (one game at a time) assuming that they would consistently choose the same team as a winner within a pair on repeated trials. Participants' accuracy levels differed from chance in both the idealized condition [mean = 55.75%, SE = 0.56%; one-sample two-tailed t test, $t(41) = 10.317$, $P < 0.001$] and the actual condition [mean = 51.47%, SE = 0.59%; one-sample two-tailed t test, $t(41) = 2.494$, $P = 0.017$]. Critically, the difference between conditions was statistically significant [dependent samples two-tailed t test, $t(41) = 6.696$, $P < 0.001$]. Median response times were significantly slower for the actual condition (mean = 1,678 ms, SE = 124 ms) than for the idealized condition [mean = 1,273 ms, SE = 38 ms; dependent samples two-tailed t test, $t(41) = 3.320$, $P = 0.002$].

Experiment 3B. A replication of experiment 3 (with $n = 156$: actual condition, $n = 78$; idealized condition, $n = 78$) was conducted using seven training games per pair of teams instead of five (196 training games total). Once again, participants in the idealized condition (mean = 56.04%, SE = 0.43%) outperformed those in the actual condition (mean = 53.27%, SE = 0.42%) at test [dependent samples two-tailed t test, $t(77) = 5.812$, $P < 0.001$], thus confirming the strong advantage produced by idealization when using real-world stimulus sets.

Cognitive Modeling. ***Recency analyses.*** Trials with extreme response times (under 150 ms or over 10 s) were eliminated from the analyses. Fig. S1 shows mean recency scores as a function of distance (D0 to D5). The average individual recency score over the studied range was positive [mean = 0.175, SE = 0.02; one-sample two-tailed t test, $t(87) = 10.533$, $P < 0.001$]. Mean scores decreased monotonically as distance increased (D0, mean = 0.338, SE = 0.032; D1, mean = 0.267, SE = 0.023; D2, mean = 0.213, SE = 0.028; D3, mean = 0.086, SE = 0.03; D4, mean = 0.032; SE = 0.032; D5, mean = −0.078, SE = 0.044), as evidenced by a

significant negative linear trend [linear trend analysis, $F(1,87) = 97.027$, $P < 0.001$].

Support vector machine models. We simulated all support vector machine (SVM) models using the scikits.learn Python-based package. For experiments 1 and 2, we simulated the SVM-Optimal model on the training data, using a Gaussian radial-basis kernel with the C parameter set at 1 (default value). ε was set to 0.5, a value that maximizes both training and test performance for all conditions. As Fig. 5 shows, test performance (defined as the proportion of trials in accord with the previously defined optimal classifier) was perfect for all conditions in experiments 1 and 2. For experiment 3, individual simulations were run using the same training and test sets as the participants. After systematic explorations, we determined that a C value very close to 0 ($C = 0.000001$) and the s parameter set at 0.28 yielded the highest performance in both conditions. As Fig. 5 shows, test performance was nearly equal for both conditions (actual condition, mean = 0.5611; idealized condition, mean = 0.5607).

SVM-Sampling simulations used the same procedures and parameters as those for the SVM-Optimal model, except for the addition of a free decision parameter γ. For experiments 1 and 2, a search for the best-fitting parameter γ value yielded a value of $\gamma = 2.12$ (mean squared error = 0.0019), whereas for experiment 3 the best-fitting value was $\gamma = 1.68$ (mean squared error = 0.0001). Fig. 5 shows that using these values to simulate the modified models leads to higher performance in the idealized condition than in the actual condition for all experiments.

Diffusion model. The EZ diffusion model (1) was fit to the choice and response time data. The EZ diffusion model was chosen because of its simplicity, ease of application, and appropriateness for our studies (e.g., it is reasonable to assume symmetrical decision boundaries). Table S2 and Fig. 6 show the results for all experiments. In all three experiments, the mean drift rate for the idealized condition was significantly higher than that of the actual condition [experiment 1, independent samples two-tailed t test, $t(84) = 9.949$, $P < 0.001$; experiment 2, independent samples two-tailed t test, $t(91) = 14.68$, $P < 0.001$; experiment 3, dependent samples two-tailed t test, $t(41) = 15.338$, $P < 0.001$]. Also, for all three experiments, the values of the boundary positions did not differ between the idealized and actual conditions [experiment 1, independent samples two-tailed t test, $t(84) = 0.602$, $P = 0.549$; experiment 2, independent samples two-tailed t test, $t(91) = 1.042$, $P = 0.300$; experiment 3, dependent samples two-tailed t test, $t(41) = 0.463$, $P = 0.646$].

1. Wagenmakers E-J, van der Maas HLJ, Grasman RPPP (2007) An EZ-diffusion model for response time and accuracy. *Psychon Bull Rev* 14(1):3–22.

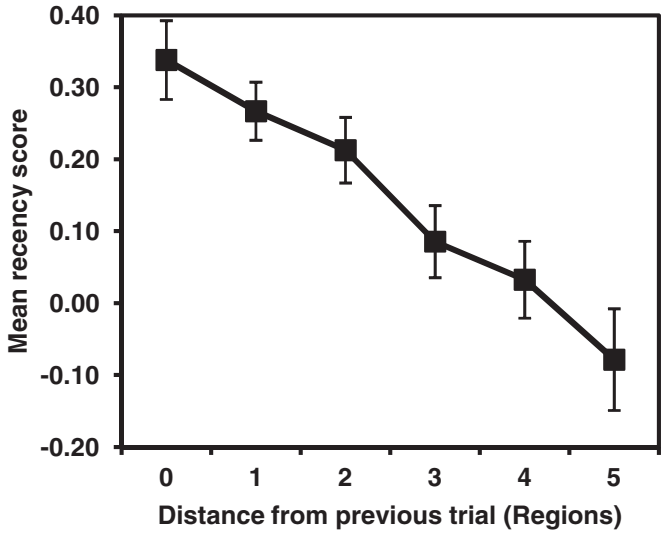

Fig. S1. Recency scores as a function of the absolute distance (in number of regions) between the stimulus for the current trial and that of the previous trial. Regions are defined in the main text. Error bars represent 95% within-subject confidence intervals.

Table S1. Measure of domain uncertainty for baseball data

	Rank for lower-ranked team						
Rank for higher-ranked team	2	3	4	5	6	7	8
1	0.548	0.714	0.667	0.738	0.881	0.810	0.857
2		0.571	0.643	0.714	0.714	0.738	0.929
3			0.429	0.738	0.762	0.810	0.738
4				0.524	0.643	0.786	0.786
5					0.476	0.667	0.833
6						0.595	0.714
7							0.643

Proportion of samples in which the higher-ranked team wins a majority of training games against the lower-ranked team. As can be seen, the domain used for experiment 3 is fairly noisy. For example, the highest-ranked team only beats the second highest-ranked team in 54.8% of the samples that were used.

Table S2. Results for the diffusion analyses (drift rates and boundary positions)

	Drift rate		Boundary position	
Experiment and condition	Mean	SE	Mean	SE
Exp. 1				
Actual	0.015	0.002	0.163	0.008
Idealized	0.1	0.009	0.156	0.007
Exp. 2				
Actual	0.014	0.003	0.158	0.007
Idealized	0.131	0.007	0.148	0.006
Exp. 3				
Actual	0.002	0.003	0.22	0.004
Idealized	0.08	0.003	0.224	0.004