A dual information seeking strategy improves imprecise human inferences outside the explore–exploit tradeoff

Yinan Cao[1,2,3,*], Clémence Alméras[1,2,3], Junseok K. Lee[1,2], and Valentin Wyart[1,2,*]

1 Département d'Études Cognitives, École Normale Supérieure, Université PSL, Paris, France.
2 Laboratoire de Neurosciences Cognitives et Computationnelles, Institut National de la Santé et de la Recherche Médicale (Inserm), Paris, France.
3 Co-first author
* Correspondence: yinan.cao@ens.fr (Y.C.), valentin.wyart@inserm.fr (V.W.)

Abstract

Everyday decisions require not just earning rewards but also learning about the world. We asked how people gather information when sampling is reward free, and compared this with reward seeking under identical outcome distributions. In a large study (N = 420), people employed two dissociable information seeking strategies. They often began by testing one option several times before switching, building early certainty, a sampling rule we call streaking. They also showed a global tendency to sample where uncertainty is greatest. Computational modeling shows that both strategies independently improve decision accuracy under noisy belief updating. Artificial neural networks trained to optimize performance acquired uncertainty-directed sampling but not early streaking, highlighting a feature of human sampling not spontaneously acquired by networks trained on these objectives. These results reveal a dual architecture of information seeking that links traits, sampling policies, and performance.

Introduction

Many everyday choices pit the safety of the familiar against the lure of the unknown [1,2]. Should we buy our favorite snack from the vending machine or risk trying something new? Psychologists and neuroscientists capture this “explore-exploit dilemma” [3] with bandit tasks, in which observers maximize cumulative reward by sequentially sampling from options with different payoff distributions [4–6]. The sequence of choices is determined by the observer’s *policy*, the systematic strategy that maps the current state of knowledge to an action. A policy may favor exploiting the best-known option or exploring to reduce uncertainty about alternatives. Typically, exploration is framed as a trade-off with immediate reward. Even with delayed outcomes, the information gathered still serves prospective valuation, that is, computing expected future payoff during sampling, much like consulting restaurant reviews before choosing where to eat [7–11].

Information itself can motivate behavior. Influential work shows that rodents and primates seek cues that reduce uncertainty about forthcoming rewards even when those cues do not alter reward contingencies [12–16]. Yet the information there still concerns the reward structure of the environment, reducing uncertainty about when or how much reward will arrive. Moreover, the informative choice is often single-shot rather than sequential, which limits what we can learn about policies that guide multi-step information gathering.

By contrast, many everyday forms of exploration are epistemic (knowledge acquisition), with policies unfolding over multiple steps. Their goal is to build internal models of the environment independently of immediate or prospective valuation. Children’s play, learning the syntax of a new language, or orienting oneself in an unfamiliar city all involve sampling information that is not tied to payoff, yet provides structural knowledge that supports flexible behavior [17–19]. How releasing sampling from prospective valuation shapes information-gathering policies is much less known.

Model-based cognitive phenotyping provides a powerful framework for probing individual variability in cognition [20,21]. A computational phenotype is defined as a set of mechanistically interpretable parameters obtained by fitting models to behavioral data. A central open question is how stable, domain-general traits relate to the latent policies that govern behavior in specific tasks. Whereas broad traits are thought to generalize across contexts [22], model-derived parameters capture the policies and computations locally recruited to solve particular problems. How these three levels—domain-general traits, task-specific policies, and behavioral performance—are related remains poorly understood. Prior work has often linked traits directly to performance, overlooking the intermediate policies that mediate between them. By applying cognitive phenotyping to epistemic exploration, we asked whether broad traits map systematically onto model-derived policies, and whether such mappings reveal hierarchical structures that explain individual differences in performance.

Here we use a two-option sequential sampling task that dissociates reward-driven from epistemic exploration under matched visual statistics. In the epistemic condition, participants adopted a distinctive policy of initial *streaking*, i.e., repeatedly sampling the same option in short streaks before switching. Although statistically suboptimal, this *tendency to streak* (ToS) co-occurred with more effective uncertainty-guided sampling, defined as a higher probability of selecting the option expected to reduce uncertainty about the hidden state, and with higher overall decision accuracy. We built a unified process model that captures the full repertoire of observed policies and formalizes streaking as a *local*, threshold-based cognitive phenotype alongside *global* uncertainty- and reward-guided policies. Critically, when belief updating is noisy, the benefit of streaking is more prevalent, consistent with an adaptive response to cognitive imprecision.

Building on this framework, individual differences revealed a hierarchical architecture. Broad personality traits predicted variation in the need for closure, which in turn related to performance *indirectly* through ToS. Finally, standard artificial neural networks, despite reproducing near-optimal uncertainty- and reward-guided sampling, failed to acquire streaking. This distinction

underscores streaking as a sampling strategy readily adopted by humans, yet not discovered by our benchmark models optimizing for task performance. It emerges through an internal drive to figure things out, testing provisional hypotheses one at a time in the face of cognitive noise. Our results support a schema in which normative global policies are complemented by local sampling rules shaped by traits and the limited precision of neural computations.

Results

Task design

Adult participants (N = 525; Methods) performed a sequential sampling task. Two hidden bags contained colored gems: a blue bag had mostly bluish gems, while an orange bag had mostly orangish gems. Each game (sequence) presented two different gems (symbols), each *independently* linked to either the blue or orange bag. On each trial, selecting a symbol revealed a shade sampled from the color distribution of its associated bag, filling the chosen symbol (Figure 1A, B). Participants did not know each symbol's source bag in advance and had to learn it through trial and error during short games of 8–20 trials.

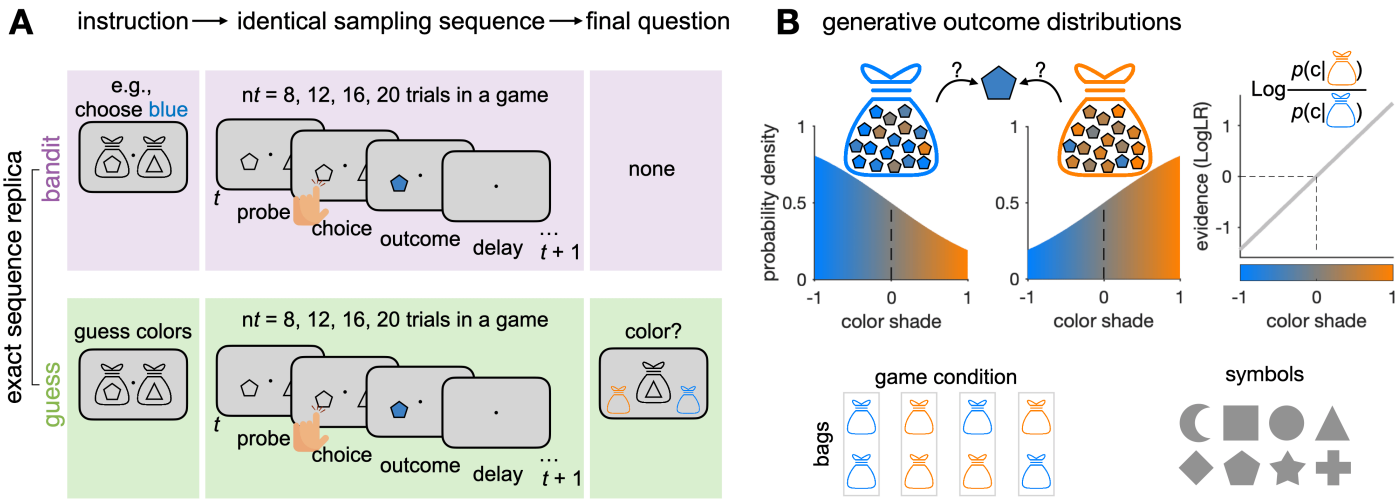


Figure 1. Experimental paradigm. (**A**) In each game, participants chose between two symbols, each *independently* associated with either a blue or an orange bag. In the bandit condition, they learned to select the symbol associated with a target color (for example, linked to the blue bag). In the guess condition, participants could freely sample the options to learn the symbol–color mappings. Both conditions included a sequential sampling phase of 8, 12, 16, or 20 choices (one choice per trial). The length of this sampling phase varied randomly across games and was not disclosed to participants before each game. In the guess condition only, participants reported after the sampling phase the color of the bag associated with a randomly probed symbol. (**B**) Outcome color shades were drawn from generative color distributions. In the blue bag, approximately two thirds of outcomes were blue and one third orange (error rate ~ 1/3), with the reverse for the orange bag. Each color shade conveyed a log-evidence (log-likelihood ratio, logLR) favoring the orange vs. blue bag, and the cumulative logLR across trials gave the posterior belief (Methods). Symbol to bag color associations were independently assigned so that both options could map to the same or to different bags.

In one condition ("bandit"), participants were instructed to choose the symbol originating from a bag of a target color (e.g., blue gems are worth more). The goal was to identify and repeatedly select the option most consistent with the target color, thereby maximizing target-aligned outcomes across trials. This instruction thus assigned value to each outcome and required participants to trade exploration against immediate reward maximization, as in standard bandit tasks. In a second condition ("guess"), there was no target color. Participants freely sampled the options to learn the mapping from symbol to bag color. At the end of each guess game, one symbol from the pair was randomly selected, and participants were asked to indicate the color of the latent bag it came from. In this condition, their performance was evaluated solely by the accuracy of this final response, with no reward provided for individual samples. This condition therefore separated exploration from any form of immediate or prospective payoff maximization.

Apart from the instructions, the two conditions were identical, allowing direct comparison of sampling patterns. Of note, a single outcome could not reveal the bag color, since each outcome was drawn from a probabilistic distribution (Figure 1B). Participants also could not infer one symbol's bag color from the other, because symbol-to-color mappings were independent: the two symbols could both map to blue, both to orange, or one to each color. After predefined quality control, 420 participants were included in the analyses (see Methods).

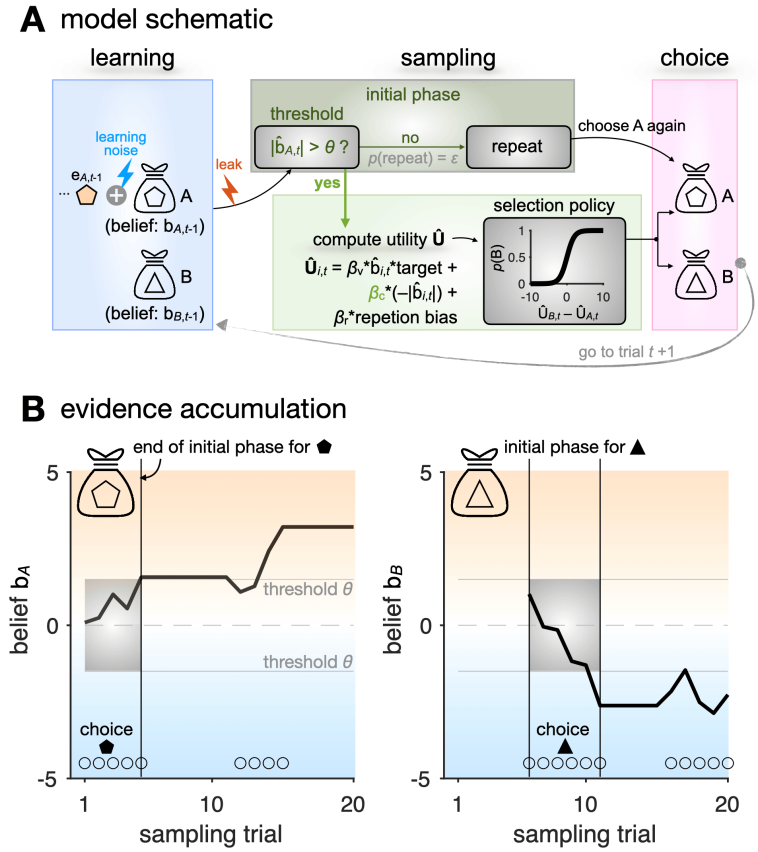


Figure 2. Computational model. (**A**) Model schematic. Learning was modeled as a noisy and leaky integration of evidence for the chosen option, with no update for the unchosen option. Choices followed a softmax over a utility that combined three components: (i) target-aligned accumulated evidence in the bandit task, weighted by a target-value sensitivity parameter (β_v); (ii) belief uncertainty, weighted by an uncertainty-sensitivity parameter (β_c); and (iii) a general repetition bias, weighted by a repetition sensitivity parameter (β_r). An initial sampling phase preceded the utility-based sampling stage. During this initial phase, choices repeated without relying on the utility until the accumulated evidence for the first sampled option crossed a threshold θ. Below this threshold, choices simply repeated with probability ε. (**B**) Evidence accumulation. For an example sequence, cumulative evidence for each option is plotted, equal to the posterior belief for each category. Only the chosen option updates. A threshold (θ of 1.5) marks the end of initial sampling for an option A, reached when its belief crosses it.

Computational model specification

Previous work with this task showed that, despite identical generative statistics, the guess and bandit conditions elicited distinct choice patterns [23]. Two forms of information seeking emerged. First, a *global* policy of directed exploration: in the bandit condition, choices tracked the target color, whereas in the guess condition, choices preferentially targeted the more uncertain option. Second, only in the guess condition did a *local* policy emerge early in each game, producing short initial streaks of repeated choices before a switch. We hypothesized that these policies would replicate in our current larger dataset (N = 420 participants). To test this, we first formalize both policies in a computational model and compare human behavior with its predictions. The model extends our previous sequential sampling model [23] by allowing noisy and leaky belief updating during sampling [24]. We present the model first to formalize a priori hypotheses and enable strong inference [25] using human behavior.

Normative benchmark. A Bayes-optimal benchmark should accumulate evidence trial by trial as log-likelihood ratios (logLR) for each option. In the bandit condition, it selects the option most aligned with the target color (most target-like). In the guess condition, where no target is defined, it selects the option with the smallest absolute accumulated evidence (most uncertain), which maximizes the expected final accuracy when the game can terminate at any moment. These are the optimal sampling policies of an ideal observer.

Noisy-leaky process model. Our model flexibly instantiates these optimal policies with two sensitivity parameters (Fig. 2A): β_c for uncertainty-guided sampling and β_v for reward-guided sampling, with β_v fixed to 0 in the guess condition. In the limit $\beta \to +\infty$ the policies are optimal, and free β parameters make the model flexible (see Supplementary Fig. 1 for illustrative model simulations). In addition, a local, threshold-based streaking policy with level θ and repetition strength ε produces an initial phase in which the agent repeats the same choice until accumulated evidence for the sampled option exceeds θ. Together, θ and ε determine the probability of staying with the current option before switching (Fig. 2B). Finally, the model included a general repetition bias parameter (β_r) for choice autocorrelation beyond the initial streaking phase, a leak parameter (δ) for gradual information loss across trials, and a learning-noise term (σ) for imprecise belief updates. These extensions capture the leaky and noisy nature of human belief updating [24,26]. Parameter and model recovery confirmed that the full model and its components were identifiable [27,28] (Supplementary Figs. S2–S3).

Next, we analyzed participants' behavior to characterize global and local information sampling in the current dataset (N = 420). We first conducted model-free analyses of choice patterns across conditions, then fit the computational model to test whether it reproduced these behaviors, and finally quantified individual differences in the policy-related parameters.

Behavioral performance and information sampling

We first assessed participants' performance in each condition. As expected, in the bandit condition, participants reliably selected the target-colored option by the final trial of each sequence (mean final trial accuracy 0.786, 95% CI [0.770, 0.802]), significantly above chance, $t(419) = 34.52$, $p < 0.0001$. Performance further improved with sequence length, $t(419) = 2.33$, $p = 0.021$, with an average slope of 0.0025 [0.0004, 0.0045]. In the guess condition, participants correctly identified the option–color mapping above chance at the final probe (mean accuracy: 0.785 [0.774, 0.797]), $t(419) = 48.96$, $p < 0.0001$, and accuracy showed a clearer increase with sequence length, $t(419) = 12.34$, $p < 0.0001$, mean slope 0.0104 [0.0088, 0.0121] (Fig. 3A).

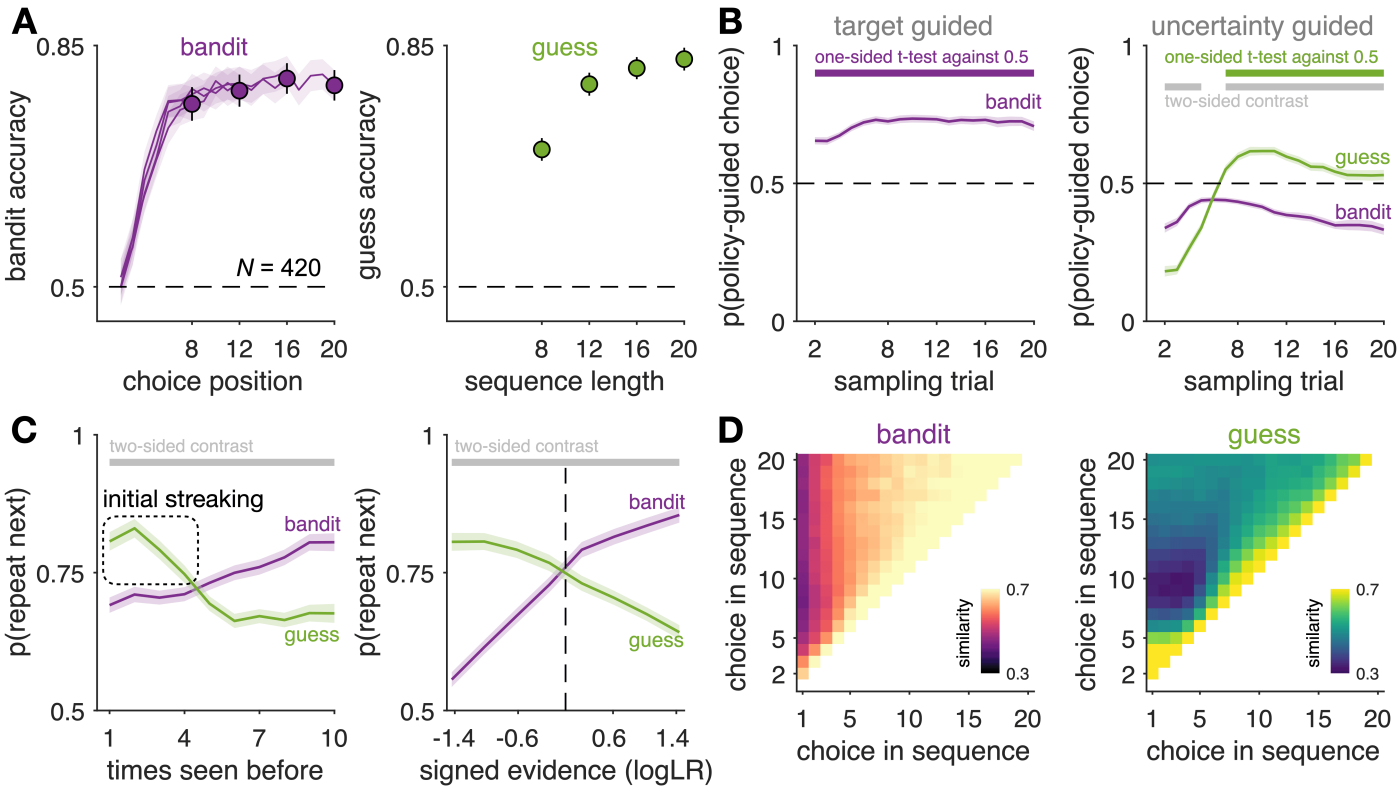


Figure 3. Human behavior. (**A**) Accuracy in the two task conditions. In the bandit condition, accuracy is the probability of selecting the option associated with the target color in games with counterfactual hidden structure (one symbol linked to orange, the other to blue). In the guess condition, accuracy is the correctness of the final probe after each game. (**B**) Fraction of choices aligned with model policies across sampling trials. On each trial, based on the outcomes observed so far, we identified the target-like option as the one with the highest accumulated evidence toward the target color, and the most uncertain option as the one with the lowest absolute accumulated evidence. Choices were counted as target-aligned in the bandit condition and as uncertainty-guided in both conditions. Thick lines indicate statistically significant effects (one-sided t-test, FDR < 0.005) and between-condition contrasts (two-sided t-test, FDR < 0.005). (**C**) Left: probability of repeating a choice as a function of how many times the current option has already been sampled. Right: probability of repeating a choice as a function of signed evidence, expressed in logLR units. In the bandit condition, signed evidence reflects target-aligned evidence; in the guess condition, it reflects belief-aligned evidence, assuming an ideal observer integrating trial-by-trial evidence for each option given the actual choices. Error bars show 95% confidence intervals across 420 participants. (**D**) Mean choice-to-choice similarity within the sampling phase, pooled across sequences of lengths 8, 12, 16, and 20.

We next tested whether participants' sampling behavior was driven by reward and uncertainty, consistent with our task design. On each trial, we classified the objectively most target-like option (highest accumulated value) and the most uncertain option (lowest absolute accumulated value) based on the running estimate of color evidence. As predicted by the respective policies, choices in the bandit condition predominantly favored the target-like option (one-sample t-test against 0.5 across trials, peak $t(419) = 35.19$, $p < 0.0001$), indicating reward-guided sampling. By contrast, choices in the guess condition showed a significantly stronger bias toward the most uncertain option (paired t-test between conditions, trials 7−20, FDR-corrected $p < 0.005$; peak $t(419) = 20.51$, $p < 0.0001$). These results confirm that participants adopted a reward-guided policy in the bandit task and an uncertainty-guided sampling policy in the guess task (Fig. 3B).

Closer inspection, however, revealed a systematic deviation from the optimal policy early in the guess sequences (around trials 2–5), where participants showed a temporary bias toward the relatively more certain option. This early behavior suggested an initial, local form of information seeking distinct from the overall uncertainty-guided policy. We quantified this as an initial *streaking* policy, a tendency to repeat choices in "streaks" (ToS) early in the game, by analyzing the fraction of repeat decisions as a function of prior samples from that option.

In the guess condition, participants showed a stronger tendency to resample options they had sampled only once or twice ($t(419) > 17.17$, $p < 0.0001$; Fig. 3C, left), revealing a strong repetition bias early in the game. This pattern was absent in the bandit condition, where participants tended to progressively converge on a preferred option. That is, in the guess condition, where participants were encouraged to learn both options' colors, they were more likely to repeatedly sample the same option upon first seeing it. This pattern was confirmed using choice-to-choice similarity matrices, which measured how similar each choice was to later choices within the same game (Fig. 3D). In the guess condition, participants showed high choice similarity among early trials (e.g., trials 1–5), followed by a shift toward the other option (e.g., trials 6–10), resulting in low similarity between these two streaks of choices. This choice pattern—resembling an AAAAABBBBB sequence—reflects an initial phase of repeated sampling of the same option followed by a switch. In addition, elevated similarity between adjacent trials across the matrix reveals a general repetition bias throughout the sequence, visible as a bright band just above the diagonal (Fig. 3D).

Complementary analyses further clarified the decision dynamics driving this sampling behavior. In the guess condition, participants were more likely to resample an option when the current outcome *conflicted* with their emerging belief about its hidden color (Fig. 3C, right), and only switched after receiving confirmatory evidence (Supplementary Fig. S4B). In the bandit condition, by contrast, resampling was more likely when the outcome *aligned* with the target color, consistent with reward-maximizing behavior (Fig. 3C, right; Supplementary Fig. S4A). Time-locked to the first switch within each game, we found that only in the guess condition did participants make their first switch after a large belief-confirmatory outcome (Supplementary Fig. S4B). Reaction times did not increase toward the end of the streak (Supplementary Fig. S4C). Together, these results support an evolving hypothesis-checking strategy in the guess task, where participants initially test one option and only switch once sufficient confirmatory evidence has been accrued for that option.

Model fits and parameter estimates

So far, model free analyses have shown that both global policies and an initial streaking rule appeared in human behavior. To test whether our process model reproduces and quantifies these behaviors, we fitted the model to each participant's choices (see Methods) [29]. Simulations with the best-fitting parameters reproduced the human sampling patterns (Fig. 4A). These included reward-guided sampling in the bandit condition, stronger uncertainty-guided sampling in the guess condition, and initial streaks of repeated choices in the guess condition but not in the bandit condition. The model also reproduced the divergent repetition tendencies across conditions: in the bandit condition, participants were more likely to repeat when evidence was consistent with the target color, whereas in the guess condition repetition increased under belief-violating evidence, reflecting a tendency to resample the same option in order to check surprising outcomes. Together, these simulations show that the model accounts for both global and local policies of human sampling.

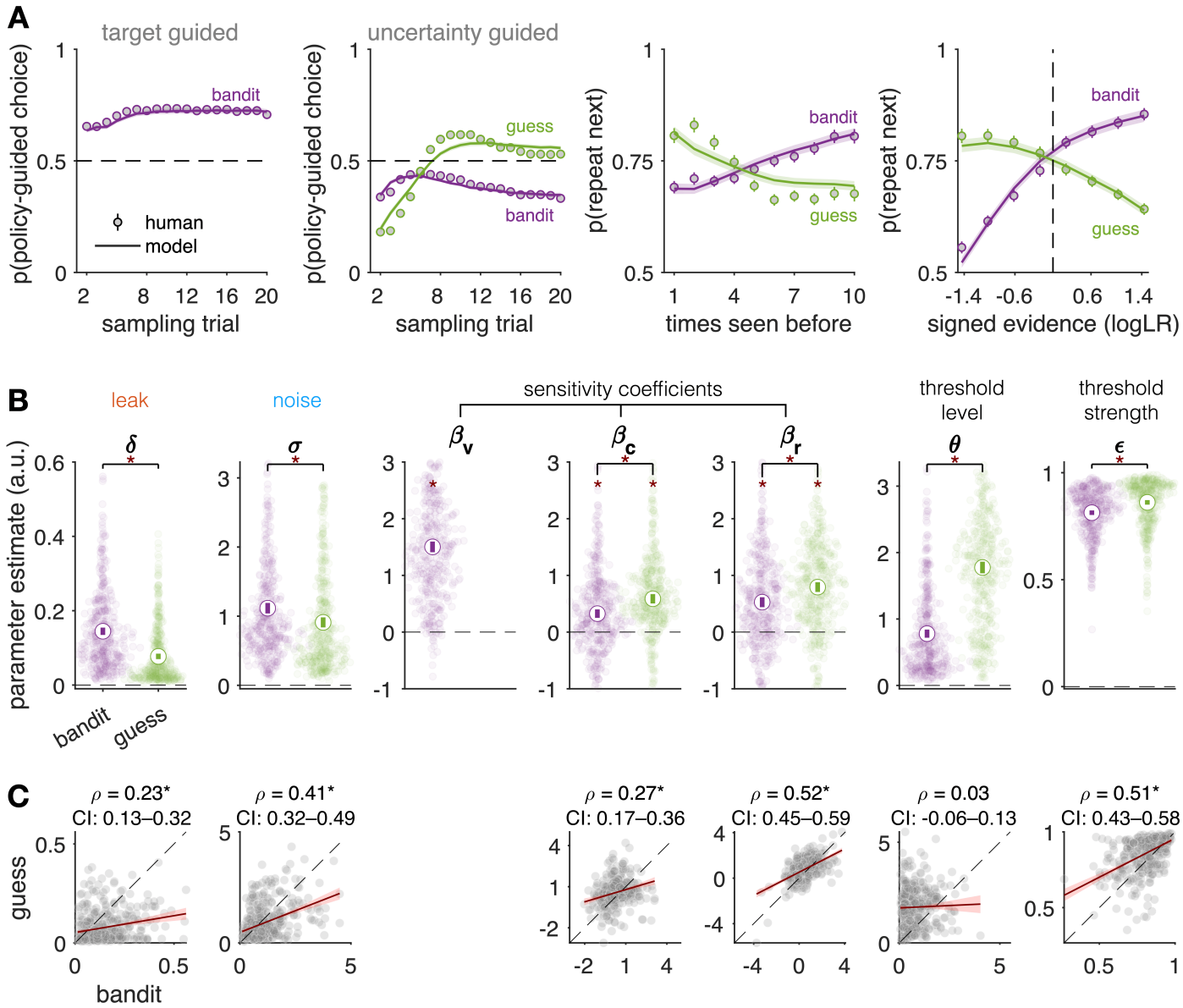


Figure 4. Model fits and parameters. (**A**) Model simulations based on the noisy-leaky belief updating model using best-fitting parameters. (**B**) Estimated parameters from the best-fitting model. Asterisks indicate statistically significant contrasts between bandit and guess conditions (paired-sample t-tests, two-sided), as well as significant deviations from zero (one-sample t-tests within conditions, two-sided), both assessed using max-statistic permutation tests with family-wise error correction (FWE < 0.05). (**C**) Parameter correlation across conditions. Asterisks indicate significant Spearman correlations between the bandit and guess conditions, assessed via bootstrapping (5,000 resamples with replacement). Bootstrap means and 95% confidence intervals are shown for each correlation estimate.

All sensitivity parameters were significantly above zero in their respective conditions ($t(419) \geq 8.51$, p values < 0.001), and they differed between conditions where defined ($|t(419)| \geq 5.01$, FWE corrected p values ≤ 0.001). In particular, uncertainty sensitivity (β_c) was markedly higher in the guess than in the bandit condition ($t(419) = 5.14$, FWE corrected $p < 0.001$), as were the threshold parameters θ and ε ($t(419) > 8.64$, FWE corrected p values ≤ 0.0002; Fig. 4B), consistent with a stronger initial streaking phase in guess. These findings corroborate that global policies governed by sensitivity parameters, and the local streaking policy governed by threshold parameters, were systematically modulated by task condition (Fig. 4B).

Learning noise σ and leak δ were lower in the guess than in the bandit. This difference suggests that shifting the task framing from reward maximization to information maximization was associated with reduced variability in belief updating and reduced decay of accumulated evidence. Thus, instructions modulated the fidelity of belief updating in addition to shaping global and local sampling policies. Across participants, parameters indexing suboptimal sampling (σ, β_r, and ε)

were positively correlated between the bandit and guess conditions (bootstrap mean $\rho > 0.4$; Fig. 4C), consistent with partially shared computational origins.

Contributions of leak and noise to suboptimal decisions

Beyond global and local sampling policies, our model fits also uncovered systematic differences in learning noise and memory leak between conditions. We therefore asked whether these imperfections quantitatively accounted for deviations from optimality. Bayesian model comparison showed that models including learning noise explained participants' sampling behavior decisively better than those without a noise term (Fig. 5A, B and D, E; family-wise comparison, noise vs. noiseless, exceedance probability $P_{exc} > 0.99$). This held in both conditions. In addition, the bandit condition featured a strong contribution of memory leak: the full model including both leak and noise best described sampling in this condition ($P_{exc} > 0.99$). By contrast, in the guess condition, learning noise emerged as the dominant source of suboptimality (Fig. 5D; Supplementary Fig. S5). Of note, randomness from the softmax rule, governed by the sensitivity parameters β, also contributed to the overall variability in choice. Learning noise captured an additional, independent source of variability beyond this selection stochasticity (see parameter recovery in Supplementary Fig. S3).

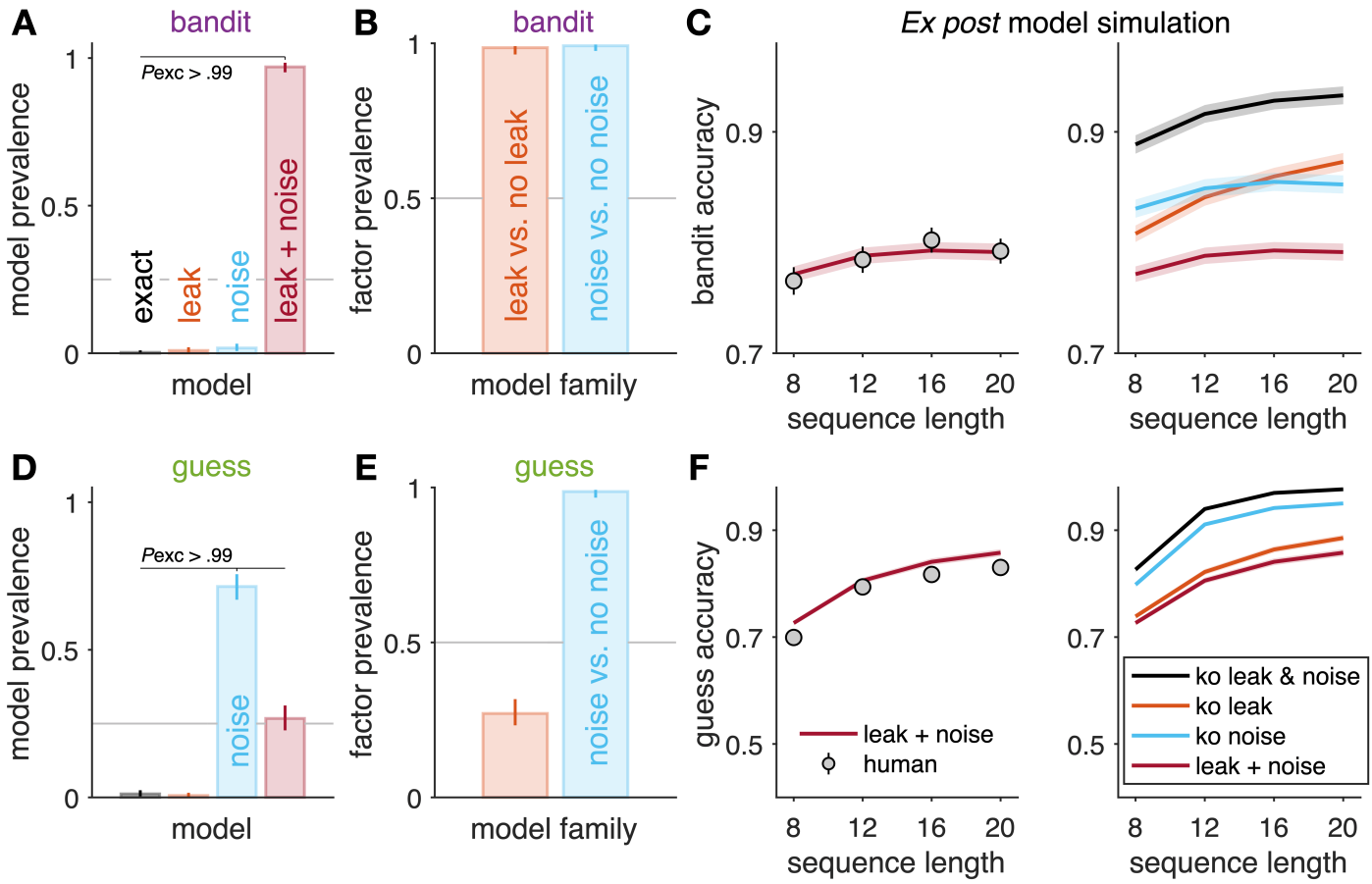


Figure 5. Model comparison and simulations. Top row: bandit condition (A–C). Bottom row: guess condition (D–F). (**A**, **D**) Posterior model frequencies from random-effects Bayesian model selection. Models: exact (no memory leak and no learning noise; other parameters retained, including sensitivity coefficients and threshold parameters), leak only, noise only, and leak + noise (full). (**B**, **E**) Family-wise comparisons testing the independent contribution of leak and learning noise to models' goodness-of-fits. (**C**, **F**) Effect on model performance of removing ('knocking out', i.e., 'ko') each factor (leak, noise, or both). Panels report sampling accuracy in the bandit task (top) and out-of-sample inference accuracy in the guess condition (bottom). Error bars indicate 95 percent confidence intervals.

To assess the behavioral impact of each factor, we simulated choice suboptimality attributable to leak and noise by "knocking out" each factor in turn, setting it to zero while holding other parameters at their best-fitting values (Fig. 5C, F). In the bandit condition, both leak and noise contributed comparably to reduced last trial accuracy (Fig. 5C). In the guess condition, however,

removing the noise term but not the leak produced a large reduction in predicted accuracy at the final probe, deviating from the observed performance (Fig. 5F).

In sum, participants' sampling patterns deviated from the optimal agent due to condition-specific computational limits: suboptimality in bandit involved both memory leak and learning noise, while in guess it was driven primarily by learning noise. Behaviorally, both conditions showed a general repetition bias ($\beta_r > 0$), with an additional, distinctive early-phase streaking strategy observed only in the guess condition.

Recurrent neural networks as process models of human exploration

Human participants showed both near-optimal uncertainty-guided sampling and suboptimal initial streaking—a combination that seems to sit in tension. Does streaking emerge naturally from optimization objectives, or does it reflect distinctively human internal drive in the face of cognitive constraints? To address this question, we trained recurrent neural networks on identical tasks as benchmark process models that explicitly optimize performance objectives.

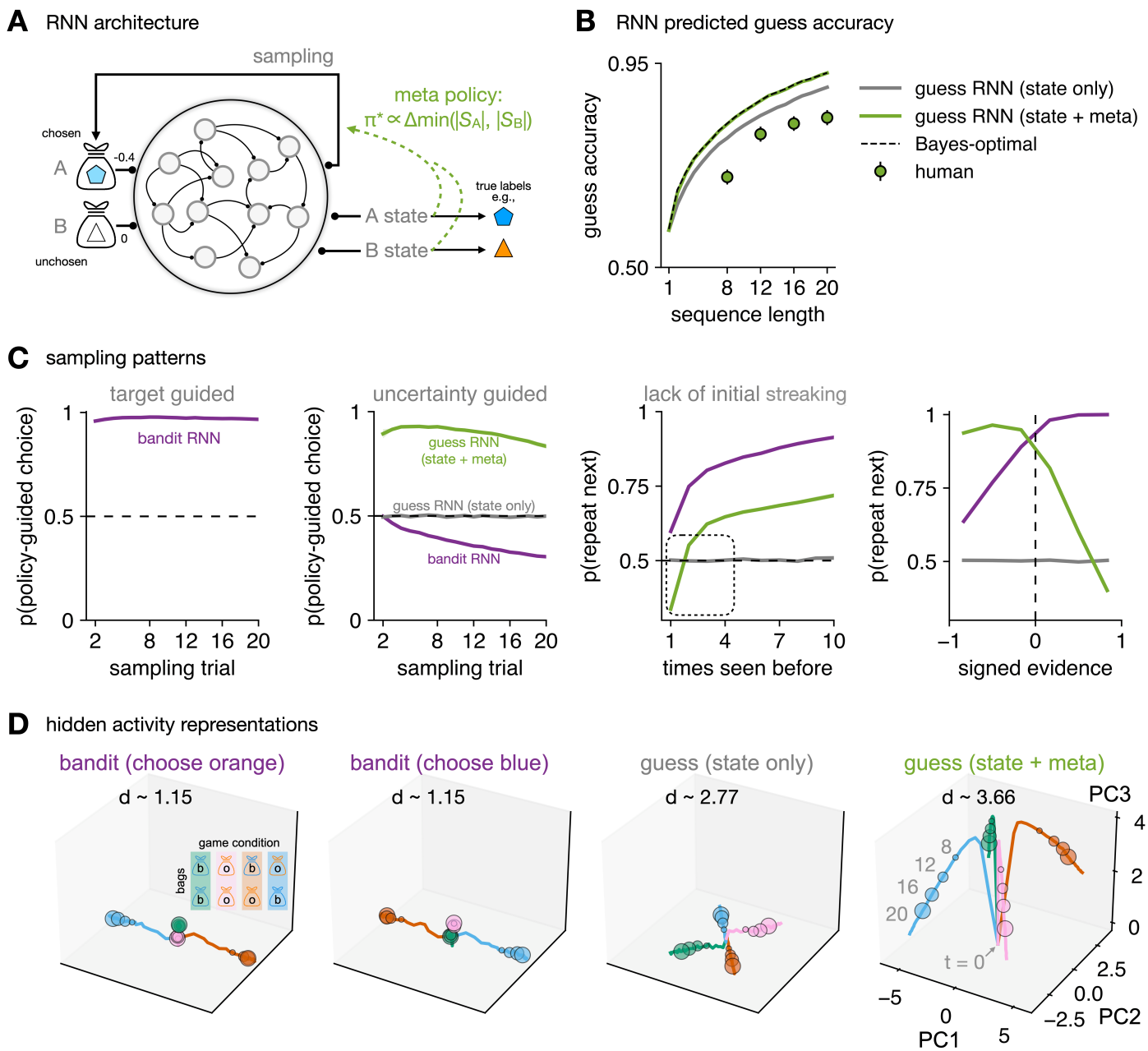


Figure 6. RNNs, sampling patterns, and hidden-activity representational geometry. (**A**) The RNN receives input conditioned on the previous sampling choice (e.g., if option A is chosen, the input is the color shade corresponding to A's outcome, while the unchosen option B yields an ambiguous zero outcome), combined with the previous recurrent activity. The updated recurrent activity is passed through a sigmoid function to produce the output sampling policy. For the bandit RNN, training relies solely on a policy-

gradient loss that maximizes cumulative reward. The reward is derived from the color shade, assigned +1 if the target is orange or −1 if the target is blue. For the guess RNN, the default objective is a cross-entropy loss on the two predicted states (sigmoid outputs) against the true states. The predicted states correspond to out-of-sample states after updates across the full sequence. The guess RNN can also include a meta-policy loss term: at each sampling trial t, the meta-policy promotes increases in a certainty signal from trial t to t + 1, where certainty is defined as the minimum of the absolute values of state A and state B. (**B**) Guess accuracy was computed by simulating each game, taking the RNN output on the final trial after the last update to the chosen option, converting it to a categorical response (orange if > 0, blue if < 0), and scoring a match to the true target. Bayes optimal: performance of an agent with perfect uncertainty guided sampling. (**C**) The bandit RNN exhibits strong reward-guided sampling policy. Only the guess RNN with the meta-policy displays uncertainty-guided sampling; the state-prediction-only RNN does not. None of the RNNs reproduce the initial streaking observed in humans. Nevertheless, they do capture key patterns: belief-confirming active sampling in the guess condition and target-seeking sampling in the bandit condition, consistent with human behavior. (**D**) Hidden activities of the bandit RNN collapse onto a one-dimensional axis for the four task conditions. By contrast, guess RNNs form separable representations of the four game conditions (i.e., different shape–bag mappings), and the addition of the meta-policy further enhances this separation, yielding a well-pronounced representational geometry. Curves: trajectories over sampling trials; circles: sequence lengths = 8 to 20 trials; 'd': effective dimensionality.

In the bandit condition, RNNs trained via a reward maximizing policy gradient converged on strong reward-guided sampling, closely paralleling human behavior (Fig. 6C). Principal component analysis of hidden activity showed that the network organized its representations primarily along a low-dimensional axis (Fig. 6D), with effective dimensionality around 1.15.

In the guess condition, RNNs trained solely to predict hidden states achieved high state-prediction accuracy (Fig. 6B) and formed a well separated representational geometry across game conditions (Fig. 6D), with effective dimensionality around two principal components (mean 2.770; Supplementary Fig. S6A). However, even with extended training (Supplementary Fig. S7) and adequate hidden layer capacity [30], these networks did not discover uncertainty-directed exploration (Fig. 6C; Supplementary Fig. 6B). Maximizing predictive accuracy alone was therefore insufficient for the spontaneous emergence of the optimal sampling policy on the vast landscape of possible strategies.

Notably, adding a meta policy objective to the guess RNN that encouraged reductions in representational uncertainty during sampling improved guess accuracy toward Bayes optimal (Fig. 6B) and produced human-like uncertainty-guided exploration (Fig. 6C). The meta policy was implemented as an auxiliary policy gradient loss on sampling choices. The meta RNN showed higher effective dimensionality in the hidden layer (mean 3.662, min 3.503, max 3.795). The third principal component encoded a moment to moment certainty signal in the outputs (Supplementary Fig. S6B).

However, a critical discrepancy remained. None of the RNNs, from 4 to 64 hidden units, reproduced the initial streaking policy observed in humans (Fig. 6C; Supplementary Fig. S8). The state-prediction RNN converged on random sampling ($\beta_c = 0$), even though forcing streaked sampling at test time in the trained network increased expected guess accuracy in late trials (Supplementary Fig. S9). By contrast, the meta-policy RNN learned the uncertainty-guided optimum ($\beta_c \to +\infty$). In that regime the marginal benefit of streaking was absent as streaking delayed a better (uncertainty-guided sampling) policy. Together, these results suggest that streaking competes with the optimal uncertainty-guided policy and yields the largest accuracy gains when sampling is suboptimal.

In sum, the failure of these networks to spontaneously generate streaking underscores that this behavior is not a default outcome of optimizing for these task objectives in standard recurrent architectures. Instead, it may reflect specific inductive biases or cognitive constraints in human learners that are not captured by these models. We next examined its functional consequences for human performance and its relationship to individual cognitive traits.

Initial streaking co-occurs with higher human performance

Using standard RNNs as a computational benchmark, we observed a key discrepancy: the networks failed to acquire initial streaking. To assess the functional significance of this human strategy, we next compared participants whose behavior was best explained with or without the streaking mechanism. We compared model variants with and without an initial streaking phase using model selection on the evidence lower bound (ELBO), which balances model fit against complexity [29,31]. In the bandit condition, the model including an initial sampling phase was far less prevalent than the simpler model without it ($\log_{10}BF_{10}$ = -106.9, 95% CI [-150.7, -62.5]; model with fixed θ at 0: P_{exc} > 0.99; Fig. 7A). By contrast, in the guess condition, the inclusion of an initial sampling phase markedly improved model fits for most participants (286 out of 420 participants; $\log_{10}BF_{10}$ = 753.4, 95% CI [643.4, 866.5]; model with free θ: P_{exc} > 0.99; Fig. 7E; also see Supplementary Fig. S10).

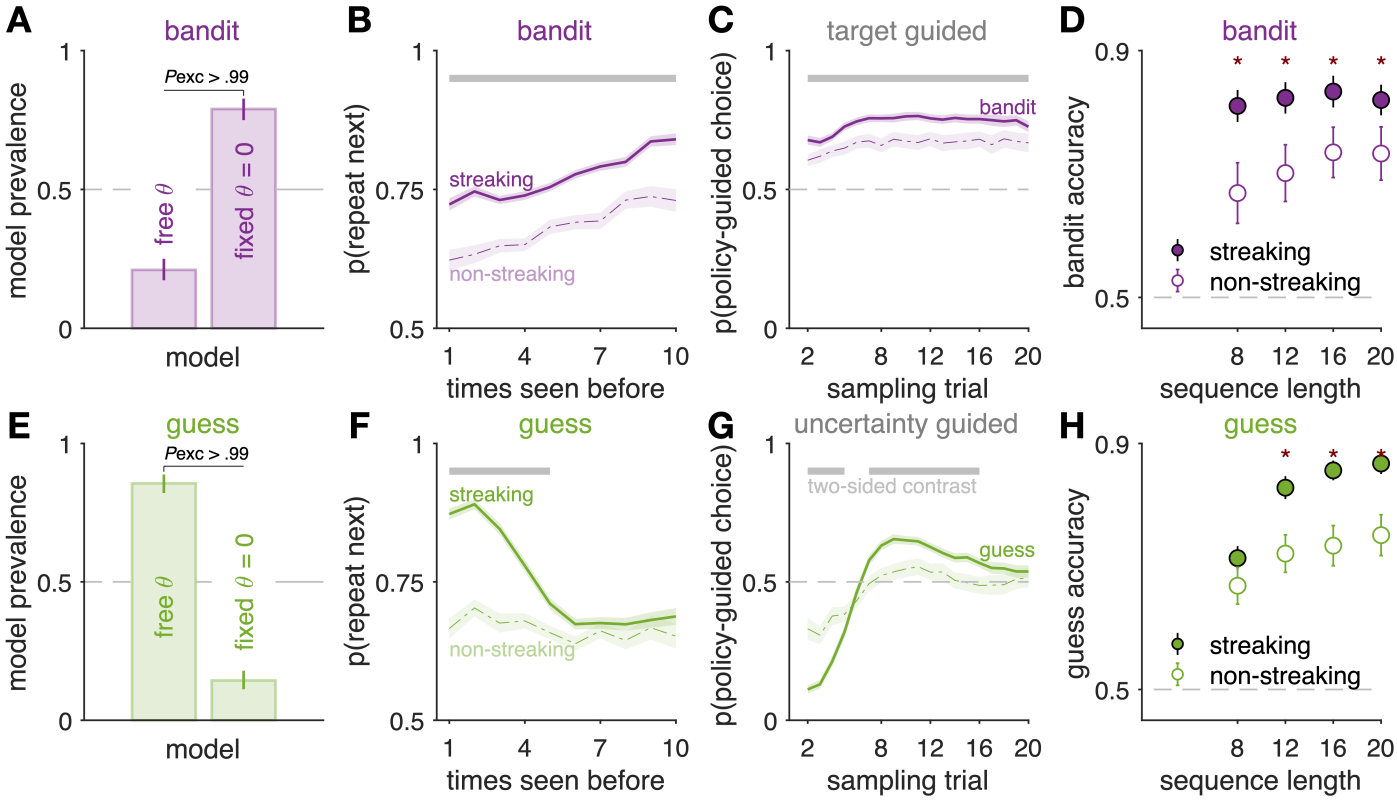


Figure 7. Distinct behaviors by initial streaking tendency. Top row: bandit condition (A–D). Bottom row: guess condition (E–H). (**A**, **E**) Individual-level model comparison between model with an initial sampling threshold and model without it, using variational ELBO as model evidence. (**B**, **C**, **F**, **G**) Fraction of choices consistent with policy-guided sampling, shown separately for participants grouped by streaking tendency. In the bandit condition, policy-guided means target-aligned choices; in the guess condition, it means uncertainty-guided choices (definitions as in Fig. 3B). (**D**, **H**) Choice accuracy for groups defined by model evidence favoring presence or absence of an initial sampling threshold. Accuracy as in Fig. 3A (bandit: target-choice rate; guess: final-probe correctness).

We divided participants into two groups based on model fits in the guess condition: those better explained by a model with an initial evidence threshold ("streaking" subgroup: N = 286) and those better explained without it ("non-streaking": N = 134). By definition, the streaking participants showed a positive tendency to streak (ToS), while non-streaking subgroup showed a negative ToS (Supplementary Fig. S10A).

As expected, streaking participants were more likely to resample options in continuous streaks during the early trials of the guess game (Fig. 7F). Later in the sequence, they also made better use of uncertainty-guided sampling, scheduling actions that best reduced uncertainty about the hidden state (Fig. 7G). In the bandit condition, streaking participants showed a stronger preference for the target-colored option, which is the optimal strategy there (Fig. 7C). Across both conditions, they also repeated their previous choice more often, although this bias did not extend to stronger

initial repetition in the bandit game (Fig. 7B). The streaking subgroup had lower leak and noise than non-streaking subgroup, but noise remained a nontrivial source of suboptimality in the streaking participants: the noisy model was decisively favored in both conditions (expected posterior family frequency > 0.98; Supplementary Fig. S10B–C).

This pattern reveals a mixture of strategies. Streaking participants combined an early suboptimal tendency to repeat with later, more adaptive directed-exploration policies. Strikingly, this combination was linked to higher performance in both conditions (bandit: $t(217.0) = 6.13$, $p < 0.0001$, $d = 0.73$; guess: $t(204.4) = 7.61$, $p < 0.0001$, $d = 0.91$; Welch's t tests, Fig. 7D, H). Thus, suboptimal repetition was not a sign of disengagement or inattention, but part of a broader strategy that, overall, improved performance.

Taken together, these results point to a distinctive role of ToS in human information seeking. Streaking improved accuracy in the guess condition, despite being suboptimal from a normative perspective. RNNs trained to optimize hidden state inference and meta cognitive uncertainty reduction did not develop this behavior even after extensive training. In this setting, streaking appears to be a distinctive feature of human information-gathering; an apparently suboptimal bias that was not reproduced by our optimization benchmarks and that nonetheless enhances performance.

Hierarchical relationships between task performance, sampling policy, cognitive ability, and psychological traits

Because sampling patterns varied widely across participants (for example, streaking vs non-streaking), we asked which cognitive abilities and psychological traits accounted for these differences. We first examined how individual sampling policies related to performance in the guess condition. ToS is a suboptimal strategy from a normative standpoint, yet it covaried with greater reliance on uncertainty guided choices. To quantify ToS as a continuous model-based index, we used model evidence favoring the initial streaking policy rather than a binary grouping.

Regression analyses showed that ToS contributed positively to guess accuracy over and above β_c and independently of the detrimental influences of leak δ and learning noise σ (Fig. 8A). Nested likelihood-ratio tests indicated that removing ToS significantly worsened regression model fit, $\chi^2(1) = 8.31$, $p = 0.004$, and removing β_c also worsened fit, $\chi^2(1) = 20.84$, $p < 0.001$, indicating that both ToS and β_c make unique, positive contributions to accuracy.

The benefit of ToS was most pronounced in late trials of longer sequences, that is, more than 12 choices. Using accuracy predicted by the full leaky-noisy model as the dependent variable reproduced the same pattern (Fig. 8A; Supplementary Table 1). By contrast, in an exact model without learning noise *during* sampling, where noise was applied only as selection variability at the final guess *after* sampling, ToS no longer improved accuracy and only β_c predicted performance (Fig. 8A; Supplementary Fig. S11). Eliminating learning noise, but not leak, eliminated the streaking benefit on simulated accuracy (Supplementary Fig. S12; see also Supplementary Fig. S13, where higher learning noise spreads the benefit across more trials). Together, these results indicate that streaking reinstates structured sampling and enhances expected decision accuracy, particularly in the presence of noise during belief updating.

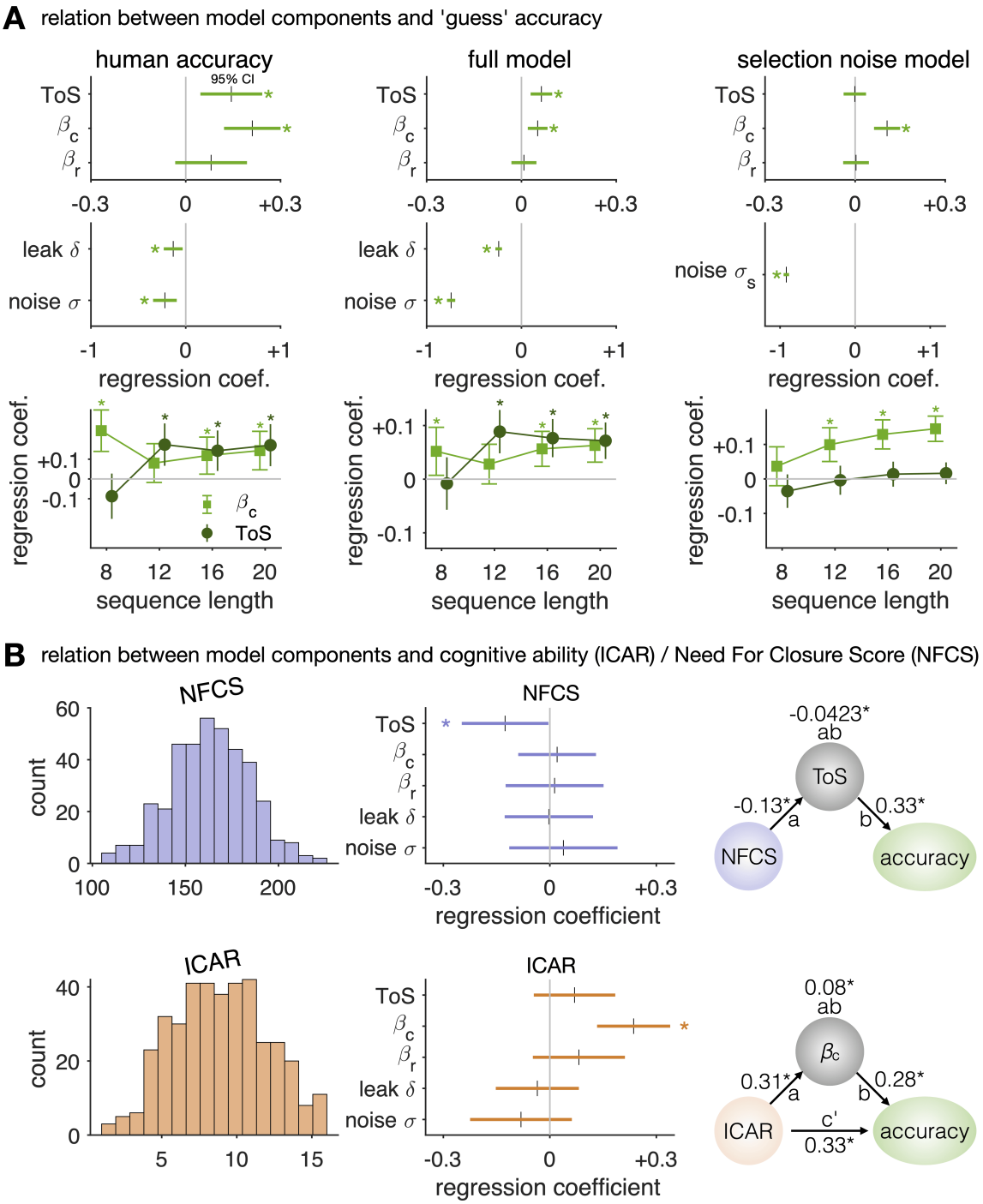


Figure 8. Relationships between performance, policy, cognitive ability, and psychological traits. (**A**) Inference accuracy in the guess condition was jointly predicted by β_c (uncertainty-directed exploration) and ToS (initial streaking-based information seeking), controlling for other model parameters. Ex post simulations with a leak + noise observer (full model) reproduced these regression effects on model-predicted accuracy. In a control model with selection noise but no learning noise, only β_c positively predicted simulated accuracy, while ToS was not significant. Top panels: mean coefficient is shown by the thin black vertical line; the thick green horizontal bar denotes the 95% CI; asterisks indicate significance against zero (two-sided one-sample t test). Bottom panels: regression analysis individually for each sequence length. (**B**) Left: histograms of questionnaire scores; middle: hypothesis-driven tests of selective associations (black vertical line: mean; horizontal bar: 95% CI); right: mediation analyses (also see Supplementary Tables 1–2).

Given the central role of ToS and β_c in supporting better accuracy, we asked how these components relate to broader dispositions. We used a top-down framework with a priori hypotheses for two established scales collected in our online sample (Fig. 8B): The Need for Cognitive Closure Scale (NFCS) and the International Cognitive Ability Resource (ICAR). NFCS indexes preference for a rapid conclusion, whereas ICAR indexes general cognitive ability. We hypothesized that NFCS would relate to ToS (willingness to continue sampling before reaching a

conclusion), and that ICAR would relate to β_c, as effective uncertainty-guided exploration requires accurately computing and tracking which option is more uncertain. Specifically, individuals with high need for closure seek rapid conclusions and may sample less extensively, whereas those with lower NFCS may be more comfortable with prolonged, structured sampling.

NFCS and ICAR were uncorrelated in our sample (Supplementary Fig. S14), indicating minimal shared variance. Multiple regression confirmed a double dissociation: only ToS significantly predicted NFCS, and only β_c significantly predicted ICAR (Fig. 8B; Supplementary Table 1). This dissociation indicates that psychological traits and cognitive ability shape distinct computational substrates of inference.

Mediation analyses clarified this hierarchy linking traits to policies to performance. NFCS did not directly predict accuracy (c' path: $p = 0.921$), but its influence was fully mediated by ToS (indirect path: $a*b = -0.0423$, 95% CI [-0.0805, -0.0052], bootstrap $p = 0.0236$; Fig. 8B). Higher NFCS scores predicted lower ToS, which in turn reduced accuracy (Supplementary Table 2). No other parameter mediated the NFCS–accuracy relationship, underscoring ToS as the selective pathway through which this dispositional trait influenced performance. This suggests that the drive for rapid closure works against the patience required for systematic, option-by-option learning, even though such structured sampling ultimately improves performance.

By contrast, ICAR predicted accuracy both directly and indirectly through β_c. Participants with higher ICAR scores showed stronger sensitivity to uncertainty, and both ICAR and β_c independently predicted higher accuracy. β_c also predicted ICAR, consistent with a reciprocal association in which cognitive ability supports adaptive inference strategies and strategy use reflects general ability. Together, these findings indicate a hierarchical architecture in which psychological traits and cognitive ability shape performance through specific latent computational processes.

Finally, in a complementary bottom-up analysis, we tested whether personality measures predicted individual differences in sampling policies. Guided by evidence that many standalone traits can be situated within the Big Five taxonomy [22], we used leave-one-participant-out cross-validation to predict NFCS, ICAR, β_c, and ToS from Big Five scores [32]. NFCS was indeed predictable ($R^2 = 0.32$; Supplementary Fig. S15; Supplementary Table 3), with positive weights for Conscientiousness and Negative Emotionality and negative weights for Open-mindedness and Extraversion. By contrast, ICAR, β_c, and ToS were only weakly predicted ($R^2 \leq 0.03$), indicating that these constructs are not well captured by broad personality dimensions. ICAR was better explained by the model parameters ($R^2 = 0.11$; see Supplementary Fig. S16–S17), consistent with the view that general cognitive ability reflects the efficiency of latent task-specific processes, while NFCS maps onto stable personality axes. This dissociation highlights the value of computational modeling: whereas psychological traits such as need for closure can be subsumed within established dispositional taxonomies, task-specific cognitive policies such as streaking and uncertainty-guided exploration reflect distinct latent processes that are not reducible to personality structure.

Discussion

We examined human information sampling when prospective valuation was removed and compared it with reward-driven sampling. We found a distinctive policy of initial streaking (ToS): upon first encountering a novel option, observers resampled it several times in a streak before switching, consistent with establishing a satisficing level of certainty. When trial-by-trial prospective valuation in payoff units was reinstated, as in the bandit condition, streaking was absent. Across individuals, stronger streaking was associated with lower need for cognitive closure and with higher decision accuracy under noisy belief updating. These findings indicate that rapid certainty buildup via streaking is an adaptive, structured policy that compensates for imprecise neural computations and improves overall decision quality.

Information sampling versus reward seeking
Neuroscience research presents contrasting views on how the brain values information. One framework treats information as intrinsically rewarding: information gain and uncertainty reduction may serve as internal reward signals [33]. Supporting this view, midbrain dopamine neurons encode non-instrumental information—cues that predict upcoming rewards without altering their probability—in ways that mirror reward prediction [12,34]. Orbitofrontal cortex similarly represents information value alongside reward [14,35]. This implies a common neural currency for evaluating both information and external rewards. However, other work emphasizes that information seeking, particularly when instrumental, can operate through distinct epistemic mechanisms not reducible to reward valuation [17,36–38]. Adjudicating between these views has been challenging because tasks typically confound informational statistics with reward structure. Our design addresses this by holding visual statistics constant across conditions while manipulating only the role of outcomes: in the guess condition, samples revealed a latent category rule instrumental for final accuracy but not expressible in reward units during sampling, whereas in the bandit condition, the same samples carried immediate payoff. This controlled contrast isolated goal-driven shifts in exploration policy, revealing that epistemic information seeking deploys distinct computational strategies beyond what reward-based frameworks predict.

The computational underpinnings of streaking
Why do participants concentrate early sampling on a single option? We reason that streaking cannot be explained by memory leak in our settings. Leak refers to the decay of accumulated evidence when an option is left unchosen. If streaking simply reflected selective maintenance, one would expect higher leak for unchosen options. Instead, the best-fitting models indicated a common leak across options, which provided a decisively better description of behavior than models allowing separate leaks. Moreover, estimated leak was lower in the guess condition than in bandit, yet streaking was stronger in guess. This dissociation rules out memory leak as the driver of streaking.

Instead, we suggest that streaking reflects a divide-and-conquer logic shaped by several cognitive constraints. First, by focusing on one option initially, the cognitive system reduces switching costs, e.g., the metabolic and computational expense of shifting attention and updating multiple belief states [39]. This is consistent with the general repetition bias ($\beta_r > 0$) we observed across both task conditions, indicating a default tendency to avoid switches.

Second, this focused sampling stabilizes state representations and sharpens coding precision for the currently active hypothesis before a switch occurs [40,41]. Initial repetitive sampling may thus reflect a cognitive schema in which participants generate a provisional hypothesis about each option before testing it through subsequent choices. Such hypothesis-testing strategies have been argued to improve efficiency in self-directed learning, since they deliver evidence perceived as more relevant at each choice [42]. Importantly, streaking should not be mistaken for simple stickiness or entrapment in a policy “saddle point”. In concept learning, it is true that participants often persist with a hypothesis until it is disconfirmed [43]. Here we find the opposite: participants continued to re-sample until evidence confirmed their current belief. They did not switch away in the face of ambiguity, but only once confirmatory evidence had accumulated (Supplementary Fig. S4B), underscoring the purposeful, confirmatory nature of the streak.

Streaking: a structured human policy not discovered by RNN learners
Our RNNs did not acquire streaking, even though they successfully learned the optimal policy of uncertainty-guided sampling when trained to maximize belief certainty. This dissociation is theoretically informative: it demonstrates that RNNs possess the computational capacity to approximate near-optimal exploration policies yet fail to spontaneously develop the local streaking rule that humans deploy.

We do not claim that streaking is impossible for any artificial network architecture. Rather, we use RNNs as a controlled computational benchmark precisely because we can systematically manipulate their optimization objective (i.e., the goal they pursue) while holding other factors constant. This mirrors our experimental manipulation in humans, where identical visual statistics yielded different exploration policies depending solely on task goals. The critical finding is the selective dissociation: RNNs trained on our tasks acquired one exploration policy (directed uncertainty sampling) but not the other (initial streaking), whereas humans exhibited both.

Streaking bears resemblance to curriculum learning effects, particularly the "blocking" advantage. Humans often exhibit a blocking advantage, learning better when related information is presented in contiguous blocks that make the underlying structure more salient. This phenomenon has been extensively studied in the curriculum learning literature [44–46]. While biologically inspired gating networks can reproduce a blocking advantage when the curriculum is externally imposed [45,47–50], these models do not actively control their own sampling. The blocking benefit in these cases arises from how the imposed input structure interacts with the network's dynamics, rather than from an emergent sampling policy chosen by the agent itself. Future work should investigate whether architectures endowed with mechanisms for self-directed sampling can discover streaking policies autonomously, thereby bridging the gap between externally imposed curriculum effects and intrinsically generated sampling policies.

Adaptive function in noisy belief updating

In our human participants, streaking benefited performance most strongly under noisy belief updating. By rapidly accumulating evidence on one option, the policy limits the dilution of updates by random fluctuations in imprecise learning [24,26,51], thereby stabilizing beliefs. In this sense, streaking illustrates how a policy that may appear locally suboptimal can be adaptive in the presence of irreducible neural noise, consistent with accounts in which structured, statistically suboptimal strategies can improve decision quality [52,53]. From a neural-coding perspective, streaking improves precision early when beliefs are weak. Neural coding precision of belief–evidence consistency is poor at low belief magnitude [54], so an early sequence such as AAAAA should yield clearer consistency decoding than an interleaved ABABA sequence at matched trial counts.

Individual differences and hierarchical architecture

Exploratory tendencies often behave like traits in humans [55] and other animals [56,57], with reliable individual differences that co-vary with broader behavioral and cognitive features, suggesting a phenotypic signature of adaptation to environmental and evolutionary pressures [58]. Here we framed individual differences at three linked levels: person-level dispositions, latent processes, and task policies. Traits such as need for cognitive closure and abilities such as ICAR index broad dispositions and capacities. Latent processes center on noisy belief updating. Policies are the strategies expressed in the task. Here, β_c reflects a global, optimal policy that allocates samples by uncertainty, whereas ToS reflects a local, bounded policy that builds early certainty by initial streaking. Empirically, ToS behaved as a policy rather than a trait, with stronger ToS associated with lower need for closure, and it co-occurred with better decision quality. β_c aligned with ICAR, consistent with a domain-general ability shaping how uncertainty is used. Together, these patterns support a hierarchical picture in which traits and abilities influence task performance through both global uncertainty-guided policy and local streaking policy.

Conclusions

Our results reveal a layered policy architecture in which an early certainty-building streaking policy emerges when reward pricing is absent, is overshadowed when it returns, and is tuned by traits and learning imprecision. Streaking highlights how stable traits may not directly map to performance in cognitive tasks, but instead shape the policies people deploy to gather information. This layered view pushes the field beyond the search for "trait correlates of exploration" and toward an understanding of how person-level dispositions tune the latent processes that govern decision policies.

Methods

Participants

The target sample size was set to 525 based on both empirical considerations and a priori power analyses. Based on our existing laboratory dataset [23], an a priori power analysis indicated that 186 participants were required to detect the smallest condition wise contrast observed previously (evidence threshold ε; $d_z = 0.18$) with 80% power at a two sided 5% significance level. No prior study has directly examined the relationship between personality measures and this task, so we set the minimal effect of interest to a correlation of $r = 0.15$. Detecting such an effect with $\alpha = 0.05$ (two-tailed) and 80% power requires 392 participants. Accounting for the typical attrition rate in online studies (~30%), we set the recruitment target at 509 participants.

Participants were recruited via the Prolific platform (www.prolific.co) and were informed about the purpose and content of the study both on the platform and immediately prior to the experiment. The study was approved by the INSERM Ethical Review Committee (IRB00003888, approval date: November 13, 2018). All participants provided informed consent electronically before beginning. They were compensated £8 for approximately 60 minutes of participation, with an additional £2 bonus contingent on good task performance. Participants were informed that the bonus would be awarded if they earned enough medals; 79 participants met this criterion and received the bonus.

Inclusion criteria. Eligible participants were required to be UK residents, fluent in English, between 18 and 35 years of age, with normal or corrected-to-normal vision (no color blindness), and to have completed at least five previous Prolific studies with an approval rate of ≥ 80%. In addition, they must not have taken part in previous laboratory studies using similar colored stimuli.

Exclusion criteria. To ensure data quality, we applied several exclusion criteria. For the model-free behavioral analyses, 105 participants were excluded, resulting in a final dataset of 420. Exclusion criteria were: (1) incomplete task performance; (2) missing trials on more than five game sequences due to technical issues; (3) production of stereotyped sequences (> 50% across all conditions), defined as game sequences where participants either chose only one option, only one response side, or systematically alternated between options or sides across the entire sequence, suggesting blind responding without outcome consideration, or (4) failure on task comprehension checks (the first four items of the debriefing questionnaire).

The retained participants' performance was comparable to that observed in the laboratory [23], indicating the reliability of the online data. For analyses linking task behavior with questionnaire responses, we further excluded 28 participants who either failed to complete the questionnaires or responded incorrectly to at least one of two catch items inserted in the NFCS and BFI-II questionnaires. In addition, one participant was excluded because all four parameter estimates (β_c, β_r, leak δ, noise σ) from the computational model fit to the guess condition were identified as outliers using Tukey's fences (interquartile range method). The final sample used for the combined behavioral and questionnaire analyses comprised 391 participants (mean age = 27.0 years, SD = 4.8; 190 female).

Task and procedure

Procedure overview. The study comprised two phases: (1) the behavioral experiment and (2) the questionnaires. Participants first received general information about the study and its procedure. After providing informed consent, they completed the behavioral task, which consisted of (a) a step-by-step tutorial explaining the rules, (b) eight practice sequences, and (c) four blocks of 16 sequences each, forming the main task (64 sequences in total, 896 sampling trials). Participants were encouraged to take self-paced breaks between blocks, but not within blocks. After the task, they could take another break before beginning the questionnaire phase, which included (a) a short debriefing on the task, (b) the Need-for-Cognitive-Closure-Scale (NFCS) questionnaire [59,60], (c) the 60-item Big Five Inventory II (BFI-II) [32], and (d) a subset of the International Cognitive

Ability Resource (ICAR) items [61,62]. Finally, participants were informed whether they had earned the £2 bonus and were invited to provide optional feedback.

Task design. Participants were instructed that they would be sampling symbols from two candidate "bags." Each bag contained only one type of symbol (e.g., one bag contained triangles, the other circles), but the individual samples could vary in color shade. On average, one bag contained more bluish shades and the other more orangish shades, though each still included about one-third of samples in the non-dominant color (Fig. 1B).

Each sampling sequence (a "game") involved two new bags. Participants did not know the underlying color distributions at the start and had to sample sequentially, one symbol per trial, from either bag. They observed only the sampled symbol's color shade, not the outcome of the unchosen bag. Sequences ended unpredictably, after 8, 12, 16, or 20 trials. Two types of games were randomly interleaved:

Bandit game. The goal was to sample symbols with the most intense shades of the target color (blue or orange, counterbalanced across participants). Each choice contributed directly to the participant's score.

Guess game. Participants sampled freely, but at the end of the sequence they were asked to identify the dominant color of one of the bags, chosen at random. Here, only the accuracy of the final response determined the score.

Participants were clearly informed at the start of each sequence which game type they were playing. They were also told that (i) the sequence could end at any time, (ii) the color distribution of one bag could not be inferred from the other (both could be blue, both be orange, or one of each), and (iii) even in a "blue" bag, some orange samples could appear, though these tended to be less intense. Thus, the bandit game emphasized immediate reward maximization, whereas the guess game required information-seeking about the hidden structure.

Trial structure. Each sequence began with a screen presenting the two bags and specifying whether the sequence was a bandit or guess game. Pressing the [space] bar initiated the trials. On each trial, the two bags appeared on either side of the screen, randomly assigned left/right, with the [f] and [j] keys used to select the left and right bags, respectively. Each trial comprised:
(a) a probe screen displaying both bags until a choice was made;
(b) an outcome screen showing the sampled symbol in its color shade for 500 ms;
(c) an inter-trial interval of 250–500 ms.

In guess games, sequences ended with a 300 ms blank screen, followed by a final question screen displaying one of the symbols and two response buttons (orange and blue; randomly assigned to the left/right sides). Participants selected its hidden bag's dominant color, saw their response highlighted (300 ms), and received feedback (the correct color for 500 ms). All sequences then ended with a 300 ms blank screen before the next instruction screen.

Feedback and incentives. The 64 sequences were divided into four blocks of 16, with equal numbers of bandit and guess games. At the end of each block, participants received feedback in the form of medals (none, bronze, silver, gold) based on their average score: (1) < 0.50: no medal (0 points); (2) 0.50–0.65: bronze (1 point); (3) 0.65–0.70: silver (2 points); (4) ≥ 0.70: gold (3 points). Participants earned the £2 bonus if they accumulated at least 6 points across the four blocks. The score was the average of game-level scores in a block: for guess games, it was the accuracy of the final questions; for bandit games, it was the fraction of choices that matched the predictions of a near-optimal reward-maximizing model with a small learning noise of 0.5. Thus, in the bandit, the incentives encouraged prospective valuation of individual sample outcomes.

Stimuli. Eight geometric symbols were combined into 28 possible symbol pairs. Colored surface size was equated across symbols. Each option (bag) was associated with a color distribution (2/3 dominant, 1/3 non-dominant). Color shades were chosen such that log-evidence for the underlying distribution scaled linearly with color intensity, ensuring that each sampled symbol carried equivalent informational value regardless of whether the dominant color was blue or orange. Outcome sequences were pre-generated, equating theoretical difficulty across conditions by simulating artificial agents.

The same 32 sequences were used in both conditions, with only the instructions differing between bandit and guess games. The 64 total sequences were arranged into four pseudo-randomized blocks of 16, balancing condition type (8 per block), bag color combinations (blue/orange, both blue, both orange), and symbol repetition.

Questionnaires. After completing the task, participants answered a set of debriefing questions and then completed two self-report questionnaires and a reasoning test, used as a proxy for IQ. The order was fixed across all participants: (1) Debriefing questions. Participants first responded to seven questions assessing their understanding of the task instructions. (2) Need for Closure Scale (NFCS) [59,60]. The NFCS is a 41-item questionnaire measuring individuals' tendency to seek certainty and avoid ambiguity. Scores up to 82 indicate a low need for closure, while scores above 205 indicate a high need for closure [60]. (3) Big Five Inventory II (BFI-II) [32]. The BFI-II is a 60-item questionnaire assessing the five broad domains of personality: Extraversion, Agreeableness, Conscientiousness, Negative Emotionality (formerly Neuroticism), and Open-Mindedness (formerly Openness to Experience). Each domain can be further divided into three facets (15 in total; Supplementary Fig. S16). A scoring key is provided by the authors. (4) Reasoning test (ICAR) [61,62]. We administered a 16-item subset of the International Cognitive Ability Resource, including 4 matrix reasoning, 4 three-dimensional rotation, 4 letter series, and 4 verbal reasoning items. Each item was multiple-choice. This subset has been validated in previous work [63] and was used here as a measure of general cognitive ability. Scoring was one point per correct response.

Computational model

Model description

The probability of generating a symbol of a given color shade c from the orange bag was defined by a sigmoid function:

$$P(c|\text{orange}) = \frac{1}{1+e^{-kc}},$$

with $c \in [-1, +1]$ (-1 = pure blue, +1 = pure orange) and slope $k > 0$ (Fig. 1B). The complementary probability was $P(c|\text{blue}) = 1 - P(c|\text{orange})$. The corresponding log-likelihood ratio is $\text{logLR} = kc$, representing the Bayesian log-evidence (e) in favor of the orange- over blue-bag hypothesis (Fig. 1B). This means that the color shades directly convey evidence about the hidden state, making the task both interpretable and tractable for modeling. The slope parameter was set to k = 1.4386, yielding an error rate of roughly one third for observing a bluish shade from the orange bag (and vice versa).

The computational model of sampling behavior included a learning module identical across both conditions. At each trial t, the model accumulated evidence $e_{i,t}$ for the chosen option i to form an evolving posterior belief $b_{i,t} = \sum_{1:t}^{i} e_{i,t} = k \sum_{1:t}^{i} c_{i,t}$. The unchosen option received no update. A utility variable was then computed as:

$$U_{i,t} = \beta_v \cdot b_{i,t} \cdot \text{target color} + \beta_c \cdot (-|b_{i,t}|) + \beta_r \cdot \text{repetion bias},$$

where β_v (bandit condition only), β_c, and β_r are sensitivity parameters to the target evidence, belief uncertainty ($-|b_{i,t}|$), and repetition bias, respectively. The "target color" was coded as −1 when the goal in the bandit condition was to choose the symbol most likely drawn from the blue bag, +1

when the target was orange, and 0 in the guess condition. The repetition bias was coded as +1 if the previous choice at trial $t-1$ was option i, and -1 otherwise, thus biasing choice to repeat on trial t when $\beta_r > 0$.

Choices were generated by one of two policies. By default, a softmax rule was applied to the utilities, with the β parameters jointly determining the effective inverse temperature,

$$p(a_t = i) = \frac{\exp(\widehat{U}_{i,t})}{\sum_j \exp(\widehat{U}_{j,t})}$$

In addition, the model allowed for an initial sampling phase: as long as the posterior belief for the chosen option remained below a threshold level θ, choices ignored utility and followed a sticky rule, repeating the previous choice with probability ε and switching with probability $1-\varepsilon$. This phase was thus governed by θ, which sets the amount of posterior belief required before switching to utility-based choice policy, and by ε, which determines the strength (hardness) of repetition under the threshold θ. To avoid discontinuities in the likelihood, θ was modeled as normally distributed with a small standard deviation of 0.1 in logLR units (notably small relative to the information provided by a single sample, up to ~1.44 logLR units).

Note that the computational model nests the optimal sampling agent for both conditions. The optimal agent does not include the initial sampling phase or the repetition bias. In the bandit condition, it selects based solely on the signed difference between the mean color-mixture estimates relative to the target color, with $\beta_v \to \infty$ (argmax choice). In the guess condition, it selects based solely on the difference in accumulated information between options, with $\beta_c \to \infty$.

Following each outcome, the posterior belief of the chosen option was updated and further corrupted by additive random learning noise drawn from a normal distribution of zero mean and standard deviation σ:

$$\tilde{b}_{\text{chosen},t} = b_{\text{chosen},t-1} + e_{\text{chosen},t-1} + \eta_t, \qquad \eta_t \sim \mathcal{N}(0, \sigma^2)$$

Moreover, a memory leak caused beliefs to drift back toward zero between trials:

$$\hat{b}_{i,t} = (1-\delta)\tilde{b}_{i,t}, \qquad 0 \le \delta \le 1$$

The leak parameter δ was shared across chosen and unchosen options. Model comparison indicated that this common-leak model provided a better fit to participants' choices than a model alternative with independent leaks for chosen and unchosen options (Supplementary Fig. S5). The final noisy-leaky utility used for choice selection was then:

$$\widehat{U}_{i,t} = \beta_v \cdot \hat{b}_{i,t} \cdot \text{target color} + \beta_c \cdot (-|\hat{b}_{i,t}|) + \beta_r \cdot \text{repetion bias}.$$

Model fitting

We estimated the conditional likelihood of each participant's decisions given the model parameters using Sequential Monte Carlo methods (particle filtering). Point estimates of best-fitting parameter values were first obtained with Bayesian Adaptive Direct Search (BADS) [64], using 10 random starting points per parameter. These point estimates then served as initializations for estimating the joint posterior distribution with Variational Bayes Monte Carlo (VBMC) [29,31]. Parameter priors used in the VBMC fitting are available in the online repository. Reported best-fitting parameters correspond to the posterior means estimated by VBMC, and model goodness-of-fit was quantified using the evidence lower bound (ELBO).

Ex post model simulations
We simulated the models in generative mode using their best-fitting parameter values (Fig. 4A; Fig. 5C, F). This procedure assessed whether, given the retrieved parameters, the models could reproduce the behavioral effects of interest without replaying participants' exact trial sequences, reward histories, or choice orders.

For the guess condition, we evaluated out-of-sample inference by simulating each game and extracting the model's posterior belief at the final trial, after the last update of the chosen option. We then applied an argmax rule to obtain the categorical response (orange if the posterior belief > 0, blue if < 0). Guess accuracy was scored by whether this response matched the true target, with chance-level performance coded as 0.5 when the posterior belief was exactly zero.

Model comparison
We performed random-effects Bayesian model selection (BMS) using the standard Dirichlet parameterization [65]. Fixed-effects comparisons were conducted using Bayes factors, by default contrasting more complex models with simpler ones (e.g., model with a free threshold vs. model without a threshold; Fig. 7A, E). For each condition, the ELBO across participants was entered into BMS to obtain exceedance probabilities along with Dirichlet-based posterior probabilities and 95% credible intervals. In addition, we ran factorial family-wise model comparisons (e.g., "leak" vs. "no leak" and "noise" vs. "no noise"; Fig. 5B, E), based on grouping models into families and estimating family exceedance probabilities under the random-effects assumption. Credible intervals for family probabilities were computed by Dirichlet sampling (10,000 draws).

Model recovery
We conducted model recovery to verify that the models under comparison were both distinguishable and recoverable within the fitting procedure (Fig. 4). Using the mean best-fitting parameters from the full model across all 420 participants and both conditions (thus avoiding condition-specific bias; except with $\beta_v = 0$ in the guess condition), we simulated data and then fit all four model variants (full model, or knocking out leak, noise, or both) to these simulations. Subsequent model comparison confirmed that each model was both identifiable and recoverable (Supplementary Fig. S2).

Parameter recovery
We performed model parameter recovery to test whether our fitting procedure introduced spurious parameter dependencies. Simulated choices were generated either from participants' best-fitting parameters (Supplementary Fig. S3A) or from random draws within predefined ranges ($\delta \in [0, 0.5]$, $\sigma \in [0, 4]$, β_v, β_c, $\beta_r \in [-3, 3]$, $\theta \in [0, 4]$, $\varepsilon \in [0.5, 1]$; Supplementary Fig. S3B). Correlations between generative parameters were destroyed by random permutation. Refitting the model showed no spurious correlations and the parameters themselves were recoverable (Supplementary Fig. S3C).

Neural network architecture
The artificial networks for the bandit and guess tasks were standard Elman RNNs. At each trial, the input X_t (color shades $\in [-1, 1]$ of options A and B, conditioned on the previous choice) and the previous hidden activity Z_{t-1} were integrated through weight matrices $\mathbf{U}$ and $\mathbf{W}$ (plus bias $\mathbf{B}$) to update the recurrent activity as in equation (1), where h_Z is a hyperbolic tangent as in equation (2). Outputs were then computed as in equation (3), where $\mathbf{V}$ are output weights. If an option was chosen, its color shade was provided as a veridical scalar value, while the unchosen option provided an ambiguous zero input. Alternatively, we also tested networks in which X_t was augmented with a two-element one-hot encoding the previous choice. This alternative architecture produced results highly consistent with those reported in the main paper.

$$\hat{Z}_t = \mathbf{W} \cdot Z_{t-1} + \mathbf{U} \cdot X_t + \mathbf{B} \tag{1}$$

$$Z_t = h_Z(\hat{Z}_t) \tag{2}$$

$$Y_t = \mathbf{V} \cdot Z_t \tag{3}$$

The output Y_t comprised three heads. The first head was passed through a sigmoid to yield action probabilities for sampling. The second and third heads were left in linear form and used only in the guess task, providing state outputs ($s_{i,t}$) analogous to posterior beliefs in the cognitive model. In the bandit task, only the first head was used. We used 64 hidden units by default and also tested smaller sizes (4, 8, 16; Supplementary Fig. S8).

Loss functions

Cross-entropy loss for state prediction

In the guess task, the RNN is only required to make a retrospective judgment at the end of each game: it has to guess which bag each symbol was drawn from. To train the RNN, we use a cross-entropy loss function to optimize the network to match its predictions to the hidden state targets (i.e., the bag colors). Importantly, the agent needs to correctly infer the hidden states of both symbols within a game; otherwise, its performance will remain at chance level. Since the network needs to infer the hidden states of both symbols, the total loss is:

$$\mathcal{L}^s(\theta) = -\sum_i y_i \log\left(p_{i,T}\right)$$

where $p_{i,T}$ denotes the out-of-sample *p*(state) output for option i, obtained by passing the corresponding output head through a sigmoid after the network has completed one episode of updates over trials $t = 0, \ldots, T$. $y \in \{0, 1\}$ is the target label for the true latent color: 0 represents blue bag, 1 orange bag. If the agent learns to predict the correct states for both symbols, it will minimize this binary cross-entropy loss.

Meta-cognitive certainty-maximization policy gradient loss

We trained the policy head to favor actions that increase a certainty signal over state representations. Let an episode have trials $t = 0, \ldots, T$. The policy $\pi_\theta(a_t|Z_t)$ outputs a Bernoulli over actions $a_t \in \{0, 1\}$ from the recurrent state Z_t. Here θ denotes the trainable RNN parameters. We define certainty signal as $\kappa_t = \min\left(|s_{1,t}|, |s_{2,t}|\right)$. The immediate reinforcement signal is the change in certainty that follows action a_t, $\Delta\kappa = \kappa_{t+1} - \kappa_t$. Action on trial $t = 0$ is forced to be random by design and excluded from learning. The objective maximizes expected certainty gain starting from $t = 1$:

$$\mathcal{J}^c(\theta) = \mathbb{E}_{\tau \sim \pi_\theta}\left[\sum_{t=1}^{T-1} \kappa_{t+1} - \kappa_t\right]$$

Using the policy gradient theorem with a one-step advantage, the loss we minimize is:

$$\mathcal{L}^c(\theta) = -\mathbb{E}\left[\sum_{t=1}^{T-1} \Delta\kappa \cdot \log \pi_\theta(a_t|Z_t)\right]$$

Reward-maximization policy gradient loss

For the bandit task, we configured the network to make a random, unbiased choice on the first trial $t = 0$. Subsequent choices were optimized with policy gradient to maximize reward accumulation. A reward r_t was delivered when the network correctly pursued a pre-defined goal, such as selecting the symbol most likely to come from the orange bag. In this formulation, the reward corresponds to the color shade outcome of the chosen option. If the target was blue, the reward was sign-flipped relative to the color shade, ensuring alignment with the target. The policy gradient loss was then defined as:

$$\mathcal{L}^r(\theta) = -\mathbb{E}\left[\sum_{t=1}^{T} r_t \cdot \log \pi_\theta(a_t|Z_t)\right]$$

Training procedure

The RNNs were parameterized by θ (the matrices $\mathbf{U}$, $\mathbf{W}$, $\mathbf{V}$, and $\mathbf{B}$ described in the "Neural network architecture" section) and trained using the REINFORCE algorithm [66]. This procedure optimizes the networks by differentiating the defined loss functions with respect to θ: $\nabla_\theta \mathcal{L}(\theta)$. In total, we trained three types of networks using different loss terms. Bandit RNN optimized $\nabla_\theta \mathcal{L}^r(\theta)$; guess RNN (state only) optimized $\nabla_\theta \mathcal{L}^s(\theta)$; while guess RNN (state + meta) optimized $\nabla_\theta[\mathcal{L}^s(\theta) + \mathcal{L}^c(\theta)]$.

For both tasks (bandit and guess), networks were trained for 50,000 epochs. Each epoch consisted of a training batch and a test batch, both generated anew. Optimization was performed with stochastic gradient ascent using the Adam optimizer and a learning rate of 0.0001. For each network type (bandit–orange, bandit–blue, guess with state prediction only, and guess with state prediction plus meta-policy), 25 RNNs were trained independently from distinct small random initializations.

Training data per epoch comprised 50 sequences of variable length (2–20 trials, drawn from a truncated geometric distribution with hazard rate 0.2). Sequence lengths were sampled until all possible lengths (2 through 20) were represented at least once. For each sequence length, 200 games were simulated across the four possible color conditions for options A/B (orange/orange, blue/blue, orange/blue, blue/orange), yielding 10,000 training games per epoch. Test data followed the same structure but with fixed sequence lengths (8, 12, 16, and 20 trials) similar to the human experiment. For both training and test, RNNs were evaluated in mini-batches, where each mini-batch grouped games of the same sequence length. For each mini-batch, cross-entropy loss, policy gradient loss, cumulative reward, and accuracy were recorded. Asymptotic performance was reached at the end of the optimization procedure in all cases (Supplementary Fig. S7).

Principal component analysis and effective dimensionality

Hidden-state dynamics were analyzed using Principal Component Analysis (PCA) to visualize population activity and assess dimensionality. First, independently trained RNN instances were aligned using Procrustes analysis to establish a common coordinate space by finding the optimal rotation and translation relative to a reference instance. A shared set of PCs was then derived for each network type from the concatenated, aligned hidden states, and the resulting trajectories were averaged for visualization. The intrinsic complexity was quantified using the effective dimension, or participation ratio, which measures the number of significantly active dimensions based on the squared singular values of the activity matrix.

Code availability

Code for computational modeling and RNNs are available at https://osf.io/q6agc/.

Acknowledgements

This work was supported by a Starting Grant from the European Research Council (ERC-StG 759341, awarded to V.W.). Y.C. was supported by a Marie Skłodowska-Curie Fellowship (101154160). All authors benefited from an institutional grant from the Agence Nationale de la Recherche (ANR-17-EURE-0017). We thank Julien Karadayi for excellent technical support with computing cluster.

Author contributions

Conceptualization, C.A. and V.W.; Methodology, C.A., V.W., and Y.C.; Software, J.K.L. and Y.C.; Validation, Y.C. and C.A.; Formal analysis, Y.C., C.A., and V.W.; Investigation and Data curation, C.A.; Writing – original draft, Y.C., C.A., and V.W.; Writing – review and editing, Y.C., C.A., V.W., and J.K.L.; Visualization, Y.C.; Supervision, V.W.; Funding acquisition, V.W.

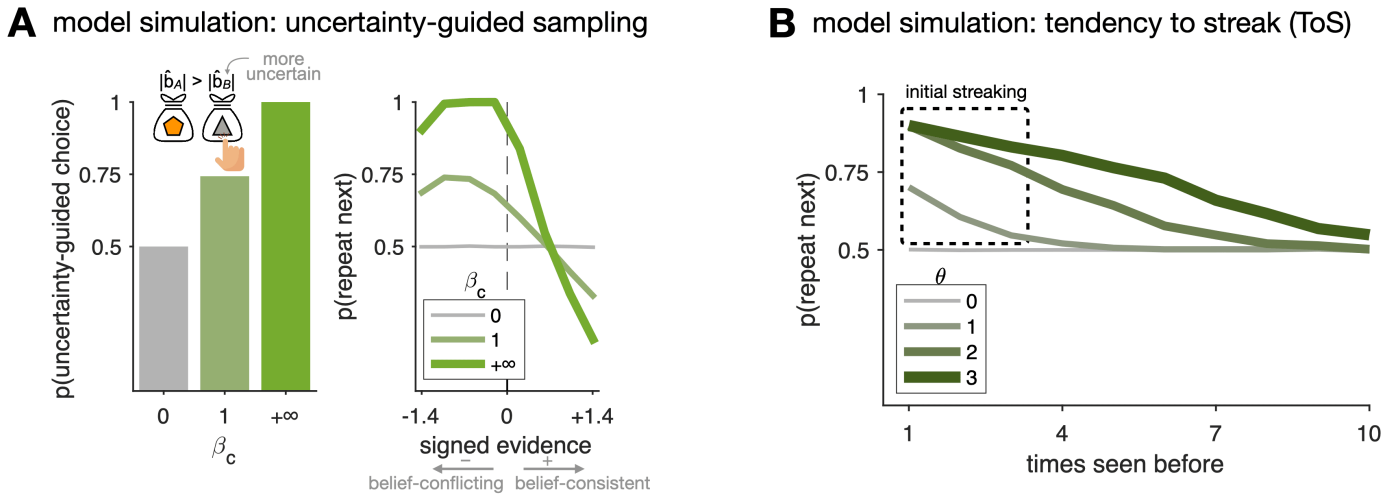
References

1. Gittins, J. C. Bandit processes and dynamic allocation indices. *J. R. Stat. Soc. Series B Stat. Methodol.* **41**, 148–164 (1979).
2. Mehlhorn, K. *et al.* Unpacking the exploration–exploitation tradeoff: A synthesis of human and animal literatures. *Decision (Wash., DC)* **2**, 191–215 (2015).
3. Sutton, R. S. & Barto, A. G. Reinforcement learning: an introduction MIT Press. *Cambridge, MA* (1998).
4. Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B. & Dolan, R. J. Cortical substrates for exploratory decisions in humans. *Nature* **441**, 876–879 (2006).
5. Cohen, J. D., McClure, S. M. & Yu, A. J. Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **362**, 933–942 (2007).
6. Schulz, E., Franklin, N. T. & Gershman, S. J. Finding structure in multi-armed bandits. *Cogn. Psychol.* **119**, 101261 (2020).
7. Wulff, D. U., Mergenthaler-Canseco, M. & Hertwig, R. A meta-analytic review of two modes of learning and the description-experience gap. *Psychol. Bull.* **144**, 140–176 (2018).
8. Wulff, D. U., Hills, T. T. & Hertwig, R. How short- and long-run aspirations impact search and choice in decisions from experience. *Cognition* **144**, 29–37 (2015).
9. Wilson, R. C., Bonawitz, E., Costa, V. D. & Ebitz, R. B. Balancing exploration and exploitation with information and randomization. *Curr Opin Behav Sci* **38**, 49–56 (2021).
10. Hills, T. T. & Hertwig, R. Information search in decisions from experience. Do our patterns of sampling foreshadow our decisions?: Do our patterns of sampling foreshadow our decisions? *Psychol. Sci.* **21**, 1787–1792 (2010).
11. Wilson, R. C., Geana, A., White, J. M., Ludvig, E. A. & Cohen, J. D. Humans use directed and random exploration to solve the explore-exploit dilemma. *J. Exp. Psychol. Gen.* **143**, 2074–2081 (2014).
12. Bromberg-Martin, E. S. & Hikosaka, O. Midbrain dopamine neurons signal preference for advance information about upcoming rewards. *Neuron* **63**, 119–126 (2009).
13. Bromberg-Martin, E. S. *et al.* A neural mechanism for conserved value computations integrating information and rewards. *Nat. Neurosci.* **27**, 159–175 (2024).
14. Blanchard, T. C., Hayden, B. Y. & Bromberg-Martin, E. S. Orbitofrontal cortex uses distinct codes for different choice attributes in decisions motivated by curiosity. *Neuron* **85**, 602–614 (2015).
15. Kang, M. J. *et al.* The wick in the candle of learning: epistemic curiosity activates reward circuitry and enhances memory: Epistemic curiosity activates reward circuitry and enhances memory. *Psychol. Sci.* **20**, 963–973 (2009).
16. Charpentier, C. J., Bromberg-Martin, E. S. & Sharot, T. Valuation of knowledge and ignorance in mesolimbic reward circuitry. *Proc. Natl. Acad. Sci. U. S. A.* **115**, E7255–E7264 (2018).
17. Kobayashi, K. & Kable, J. W. Neural mechanisms of information seeking. *Neuron* (2024) doi:10.1016/j.neuron.2024.04.008.
18. Monosov, I. E. Curiosity: primate neural circuits for novelty and information seeking. *Nat. Rev. Neurosci.* (2024) doi:10.1038/s41583-023-00784-9.
19. Gopnik, A. Childhood as a solution to explore-exploit tensions. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **375**, 20190502 (2020).
20. Patzelt, E. H., Hartley, C. A. & Gershman, S. J. Computational phenotyping: Using models to understand individual differences in personality, development, and mental illness. *Personal. Neurosci.* **1**, e18 (2018).
21. Schurr, R., Reznik, D., Hillman, H., Bhui, R. & Gershman, S. J. Dynamic computational phenotyping of human cognition. *Nat. Hum. Behav.* **8**, 917–931 (2024).
22. Bainbridge, T. F., Ludeke, S. G. & Smillie, L. D. Evaluating the Big Five as an organizing framework for commonly used psychological trait scales. *J. Pers. Soc. Psychol.* **122**, 749–777 (2022).
23. Alméras, C., Chambon, V. & Wyart, V. Competing cognitive pressures on human exploration in the absence of trade-off with exploitation. *PsyArXiv* (2022) doi:10.31234/osf.io/9qpuz.
24. Drugowitsch, J., Wyart, V., Devauchelle, A.-D. & Koechlin, E. Computational Precision of Mental Inference as Critical Source of Human Choice Suboptimality. *Neuron* **92**, 1398–1411 (2016).
25. Platt, J. R. Strong Inference: Certain systematic methods of scientific thinking may produce much more rapid progress than others. *Science* **146**, 347–353 (1964).
26. Lee, J. K., Rouault, M. & Wyart, V. Adaptive tuning of human learning and choice variability to unexpected uncertainty. *Sci Adv* **9**, eadd0501 (2023).
27. Palminteri, S., Wyart, V. & Koechlin, E. The Importance of Falsification in Computational Cognitive Modeling. *Trends Cogn. Sci.* **21**, 425–433 (2017).

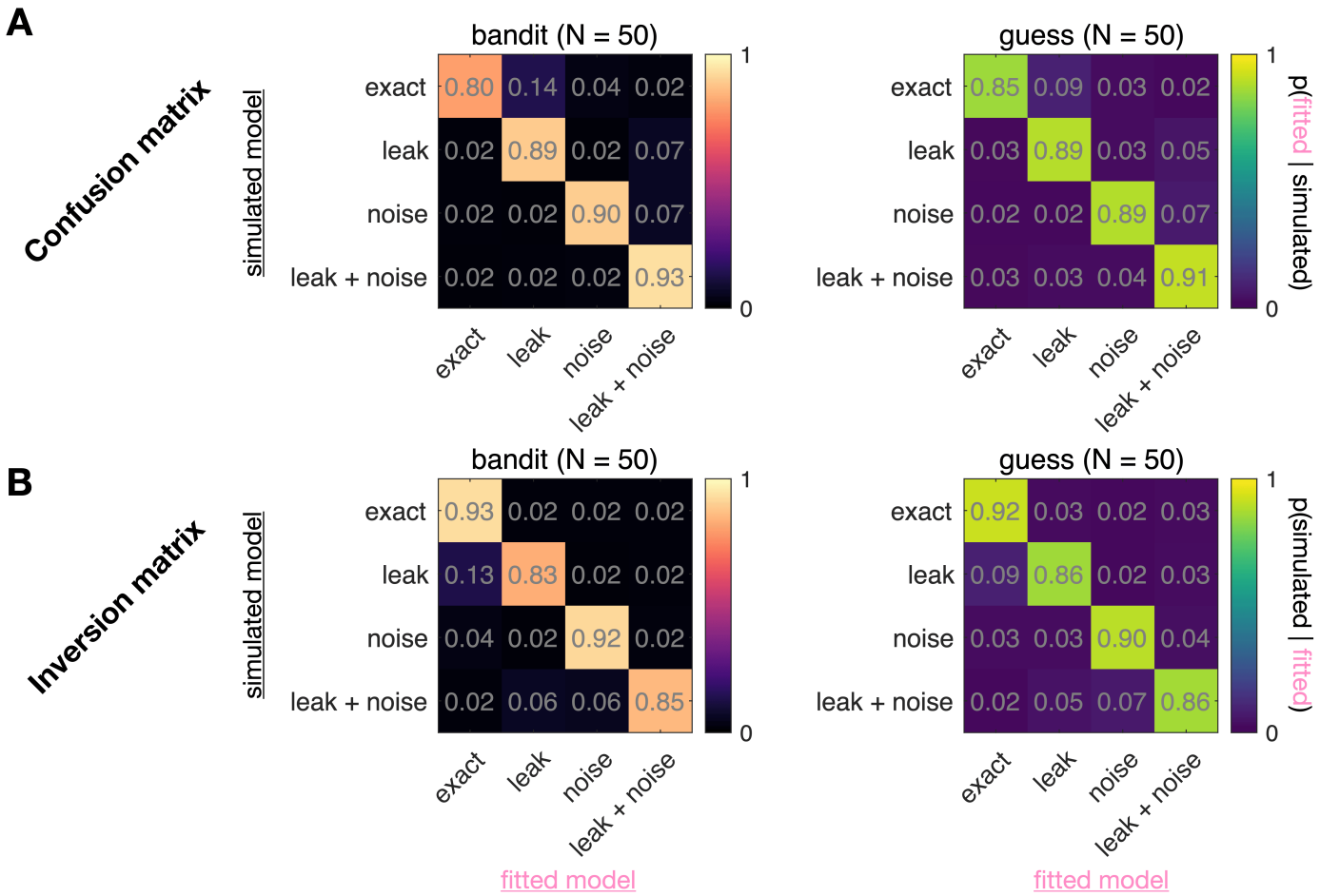
28. Wilson, R. C. & Collins, A. G. E. Ten simple rules for the computational modeling of behavioral data. *Elife* **8**, e49547 (2019).
29. Acerbi, L. Variational Bayesian Monte Carlo with noisy likelihoods. *Neural Inf Process Syst* **abs/2006.08655**, 8211–8222 (2020).
30. Collins, J., Sohl-Dickstein, J. & Sussillo, D. Capacity and trainability in recurrent neural networks. *arXiv [stat.ML]* (2016).
31. Acerbi, L. Variational Bayesian Monte Carlo. *arXiv [stat.ML]* (2018).
32. Soto, C. J. & John, O. P. The next Big Five Inventory (BFI-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *J. Pers. Soc. Psychol.* **113**, 117–143 (2017).
33. Gottlieb, J. & Oudeyer, P.-Y. Towards a neuroscience of active sampling and curiosity. *Nat. Rev. Neurosci.* **19**, 758–770 (2018).
34. Bromberg-Martin, E. S., Matsumoto, M. & Hikosaka, O. Dopamine in motivational control: rewarding, aversive, and alerting. *Neuron* **68**, 815–834 (2010).
35. Daddaoua, N., Lopes, M. & Gottlieb, J. Intrinsically motivated oculomotor exploration guided by uncertainty reduction and conditioned reinforcement in non-human primates. *Sci. Rep.* **6**, 20202 (2016).
36. FitzGerald, T. H. B., Dolan, R. J. & Friston, K. Dopamine, reward learning, and active inference. *Front. Comput. Neurosci.* **9**, 136 (2015).
37. Kidd, C. & Hayden, B. Y. The psychology and neuroscience of curiosity. *Neuron* **88**, 449–460 (2015).
38. Sharot, T. & Sunstein, C. R. How people decide what they want to know. *Nat. Hum. Behav.* **4**, 14–19 (2020).
39. Monsell, S. Task switching. *Trends Cogn. Sci.* **7**, 134–140 (2003).
40. Kool, W., McGuire, J. T., Rosen, Z. B. & Botvinick, M. M. Decision making and the avoidance of cognitive demand. *J. Exp. Psychol. Gen.* **139**, 665–682 (2010).
41. Lieder, F. & Griffiths, T. L. Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behav. Brain Sci.* **43**, e1 (2019).
42. Markant, D. B. & Gureckis, T. M. Is it better to select or to receive? Learning via active and passive hypothesis testing. *J. Exp. Psychol. Gen.* **143**, 94–122 (2014).
43. Bonawitz, E., Denison, S., Gopnik, A. & Griffiths, T. L. Win-Stay, Lose-Sample: a simple sequential algorithm for approximating Bayesian inference. *Cogn. Psychol.* **74**, 35–65 (2014).
44. Beukers, A. O. *et al.* Blocked training facilitates learning of multiple schemas. *Commun. Psychol.* **2**, 28 (2024).
45. Dekker, R. B., Otto, F. & Summerfield, C. Curriculum learning for human compositional generalization. *Proc. Natl. Acad. Sci. U. S. A.* **119**, e2205582119 (2022).
46. Collins, A. G. E. & Frank, M. J. Within- and across-trial dynamics of human EEG reveal cooperative interplay between reinforcement learning and working memory. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 2502–2507 (2018).
47. Rougier, N. P., Noelle, D. C., Braver, T. S., Cohen, J. D. & O'Reilly, R. C. Prefrontal cortex and flexible cognitive control: rules without symbols. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 7338–7343 (2005).
48. Flesch, T., Nagy, D. G., Saxe, A. & Summerfield, C. Modelling continual learning in humans with Hebbian context gating and exponentially decaying task signals. *PLoS Comput. Biol.* **19**, e1010808 (2023).
49. Giallanza, T., Campbell, D. & Cohen, J. D. Toward the emergence of intelligent control: Episodic Generalization and Optimization. *Open Mind* **8**, 688–722 (2024).
50. Russin, J., Zolfaghar, M., Park, S. A., Boorman, E. & O'Reilly, R. C. A neural network model of continual learning with cognitive control. *CogSci* **44**, 1064–1071 (2022).
51. Findling, C., Skvortsova, V., Dromnelle, R., Palminteri, S. & Wyart, V. Computational noise in reward-guided learning drives behavioral variability in volatile environments. *Nat. Neurosci.* **22**, 2066–2077 (2019).
52. Tsetsos, K. *et al.* Economic irrationality is optimal during noisy decision making. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 3102–3107 (2016).
53. Findling, C. & Wyart, V. Computation noise in human learning and decision-making: origin, impact, function. *Current Opinion in Behavioral Sciences* **38**, 124–132 (2021).
54. Weiss, A., Chambon, V., Lee, J. K., Drugowitsch, J. & Wyart, V. Interacting with volatile environments stabilizes hidden-state inference and its brain signatures. *Nat. Commun.* **12**, 2228 (2021).
55. Mkrtchian, A., Valton, V. & Roiser, J. P. Reliability of decision-making and reinforcement learning computational parameters. *Comput. Psychiatr.* **7**, 30–46 (2023).
56. Bell, A. M., Hankison, S. J. & Laskowski, K. L. The repeatability of behaviour: a meta-analysis. *Anim. Behav.* **77**, 771–783 (2009).

57. Dingemanse, N. J., Both, C., Drent, P. J., van Oers, K. & van Noordwijk, A. J. Repeatability and heritability of exploratory behaviour in great tits from the wild. *Anim. Behav.* **64**, 929–938 (2002).
58. Sih, A. & Del Giudice, M. Linking behavioural syndromes and cognition: a behavioural ecology perspective. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **367**, 2762–2772 (2012).
59. Webster, D. M. & Kruglanski, A. W. Individual differences in need for cognitive closure. *J. Pers. Soc. Psychol.* **67**, 1049–1062 (1994).
60. Roets, A. & Van Hiel, A. Separating ability from need: clarifying the dimensional structure of the Need for Closure Scale. *Pers. Soc. Psychol. Bull.* **33**, 266–280 (2007).
61. Condon, D. M. & Revelle, W. The international cognitive ability resource: Development and initial validation of a public-domain measure. *Intelligence* **43**, 52–64 (2014).
62. Dworak, E. M., Revelle, W., Doebler, P. & Condon, D. M. Using the International Cognitive Ability Resource as an open source tool to explore individual differences in cognitive ability. *Pers. Individ. Dif.* **169**, 109906 (2021).
63. Gillan, C. M., Kosinski, M., Whelan, R., Phelps, E. A. & Daw, N. D. Characterizing a psychiatric symptom dimension related to deficits in goal-directed control. *Elife* **5**, (2016).
64. Acerbi, L. & Ji, W. Practical Bayesian optimization for model fitting with Bayesian adaptive direct search. *Neural Inf Process Syst* **30**, 1836–1846 (2017).
65. Stephan, K. E., Penny, W. D., Daunizeau, J., Moran, R. J. & Friston, K. J. Bayesian model selection for group studies. *Neuroimage* **46**, 1004–1017 (2009).
66. Sutton, R., McAllester, D. A., Singh, S. & Mansour, Y. Policy gradient methods for reinforcement learning with function approximation. *Neural Inf Process Syst* **12**, 1057–1063 (1999).

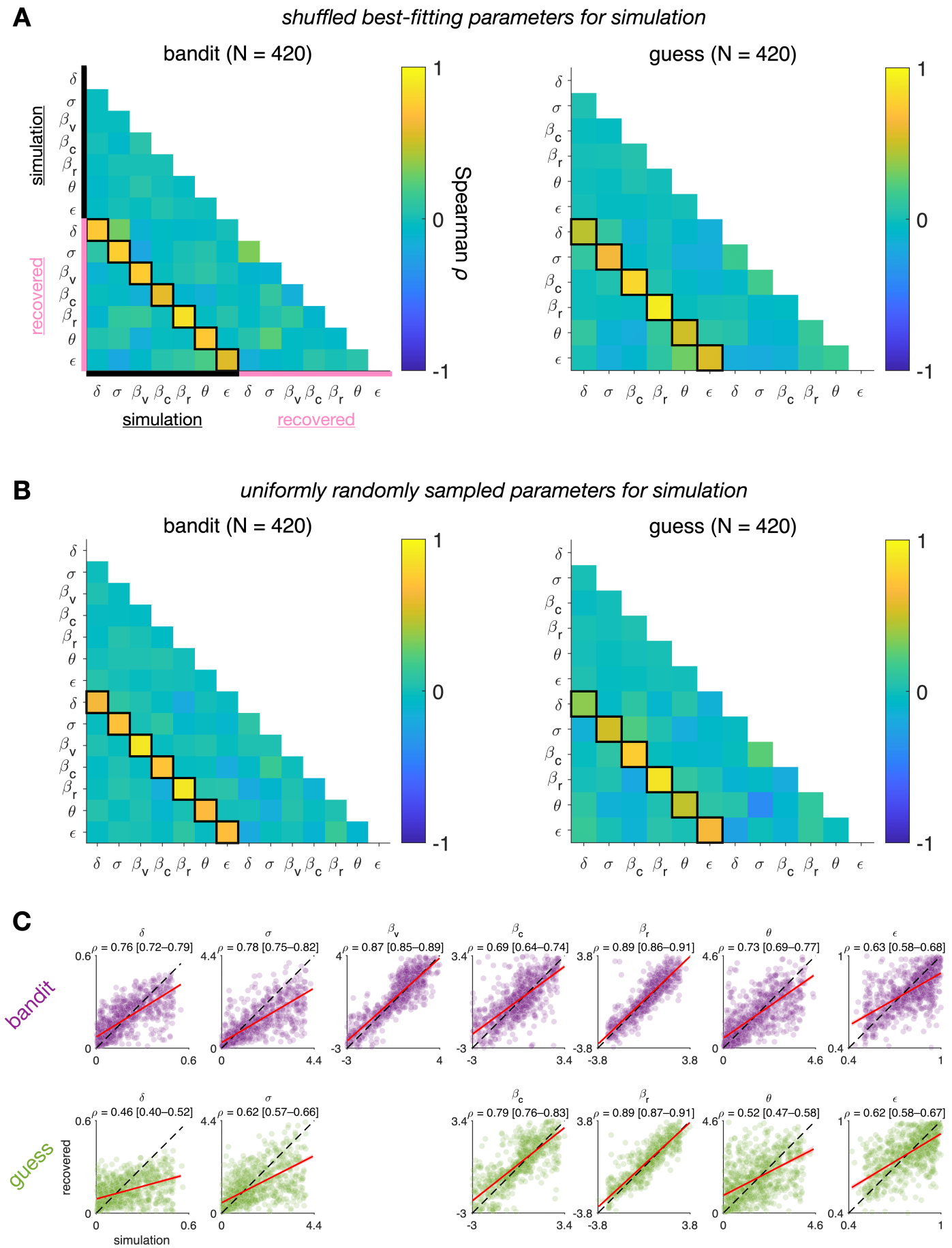
Supplementary figures



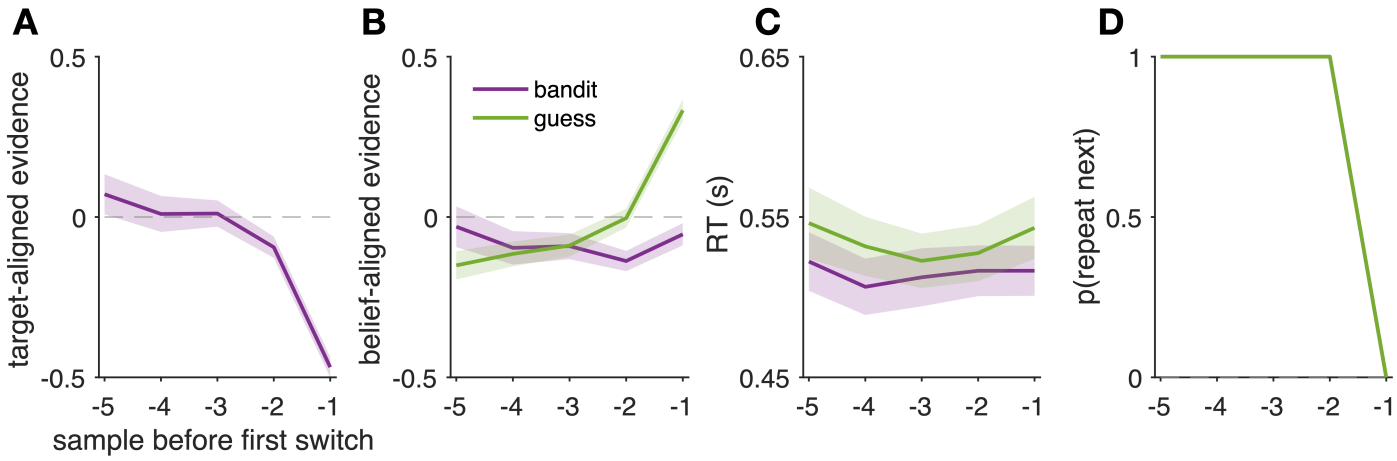
Supplementary Fig. S1. Illustrative model simulations. (**A**) (left) Varying β_c changed the tendency to direct choices toward the relatively more uncertain option. β_v, β_r and θ were set to 0; (right) A belief-aligned repetition bias emerged as a byproduct of uncertainty-guided sampling. Uncertainty is defined as the absolute cumulative logLR (belief strength) for each option. When $\beta_c > 0$ the model selects the option with the smaller belief, which reduces expected uncertainty about the symbol–color mapping. A byproduct of this policy is a tendency to repeat after belief-conflicting evidence: a modestly conflicting sample pulls belief toward zero, making the just-sampled option the more uncertain one, which leads to a repeat. Belief-consistent evidence increases certainty about the sampled option, making the other option comparatively less known and prompting a switch. (**B**) Varying the threshold level θ modulated the tendency to produce initial choice streaking. To isolate the threshold mechanism, the other sensitivity parameters (β_v, β_c, and β_r) were set to zero, repetition strength ε to 0.9 in these examples. Note that $\theta = 0$ removes the initial sampling phase, rendering ε irrelevant, since repetition under the threshold never applies. The threshold θ explicitly controls early streaking. For $\theta > 0$, the agent repeats the current option until its belief crosses θ, then switches and repeats the other (e.g., AAAA BBBB).



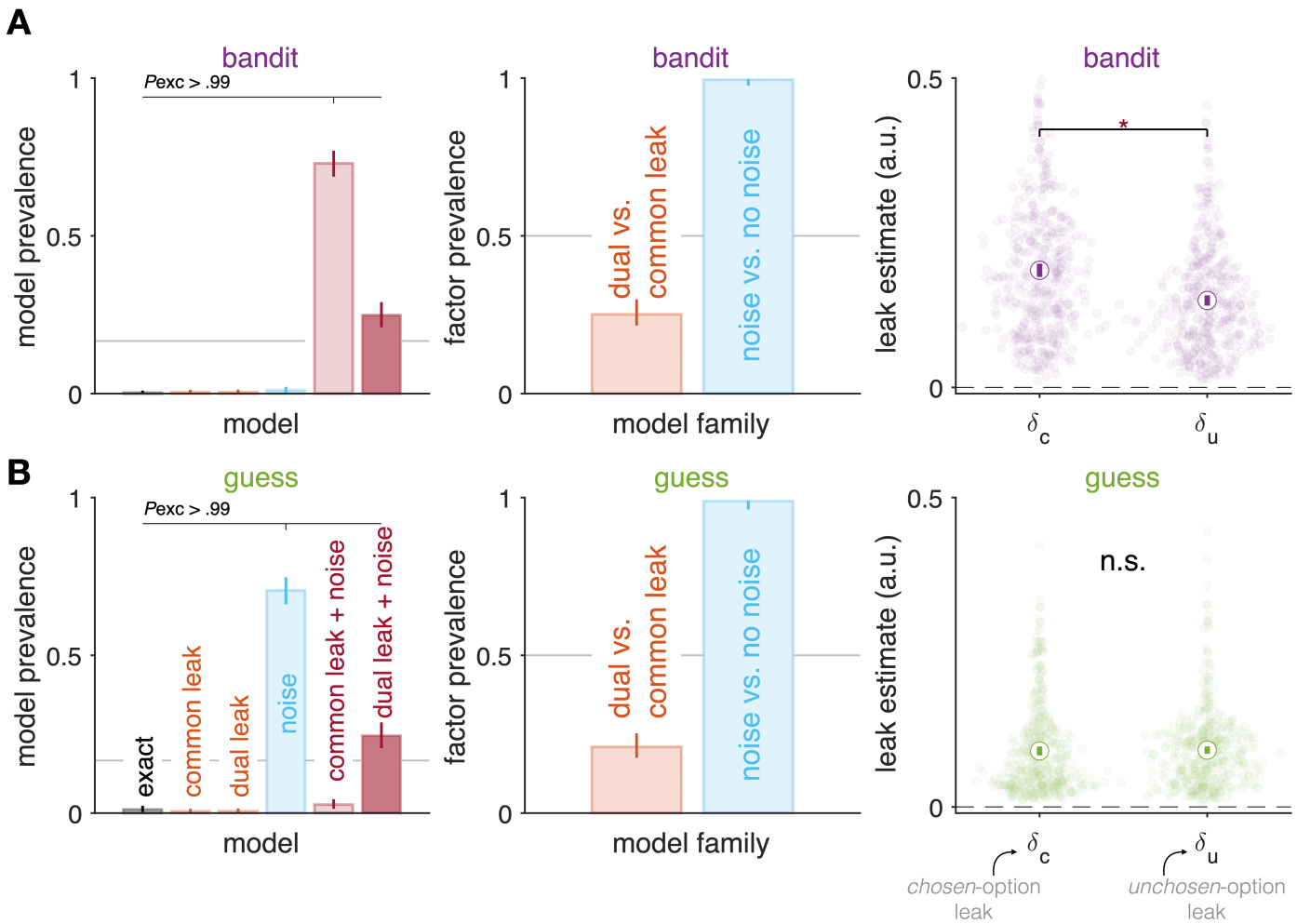
Supplementary Fig. S2. Model recovery. Using the mean best-fitting parameters from the full model across all 420 participants and both conditions (thus avoiding condition-specific bias; except with $\beta_v = 0$ in the guess condition), we simulated data and then fit all four model variants (full model, or knocking out leak, noise, or both) to these simulations. Color intensity indicates posterior model frequency.



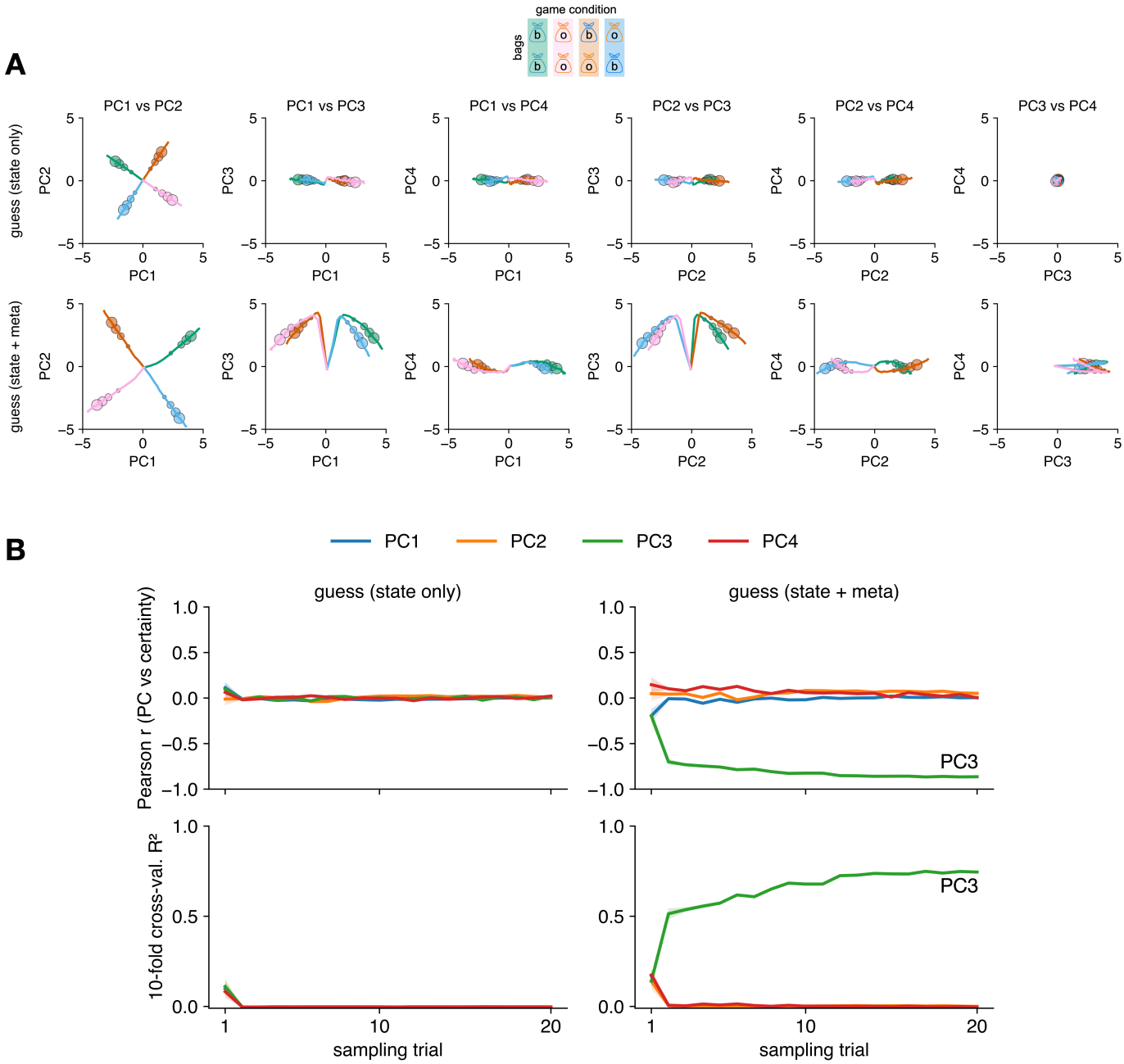
Supplementary Fig. S3. Parameter recovery. Simulated choices were generated either from participants' best-fitting parameters (A) or from random draws within predefined ranges ($\delta \in [0, 0.5]$, $\sigma \in [0, 4]$, β_v, β_c, $\beta_r \in [-3, 3]$, $\theta \in [0, 4]$, $\varepsilon \in [0.5, 1]$; B). Correlations between generative parameters were destroyed by random permutation. Refitting the model showed no spurious correlations (A, B) and the parameters themselves were recoverable (C).



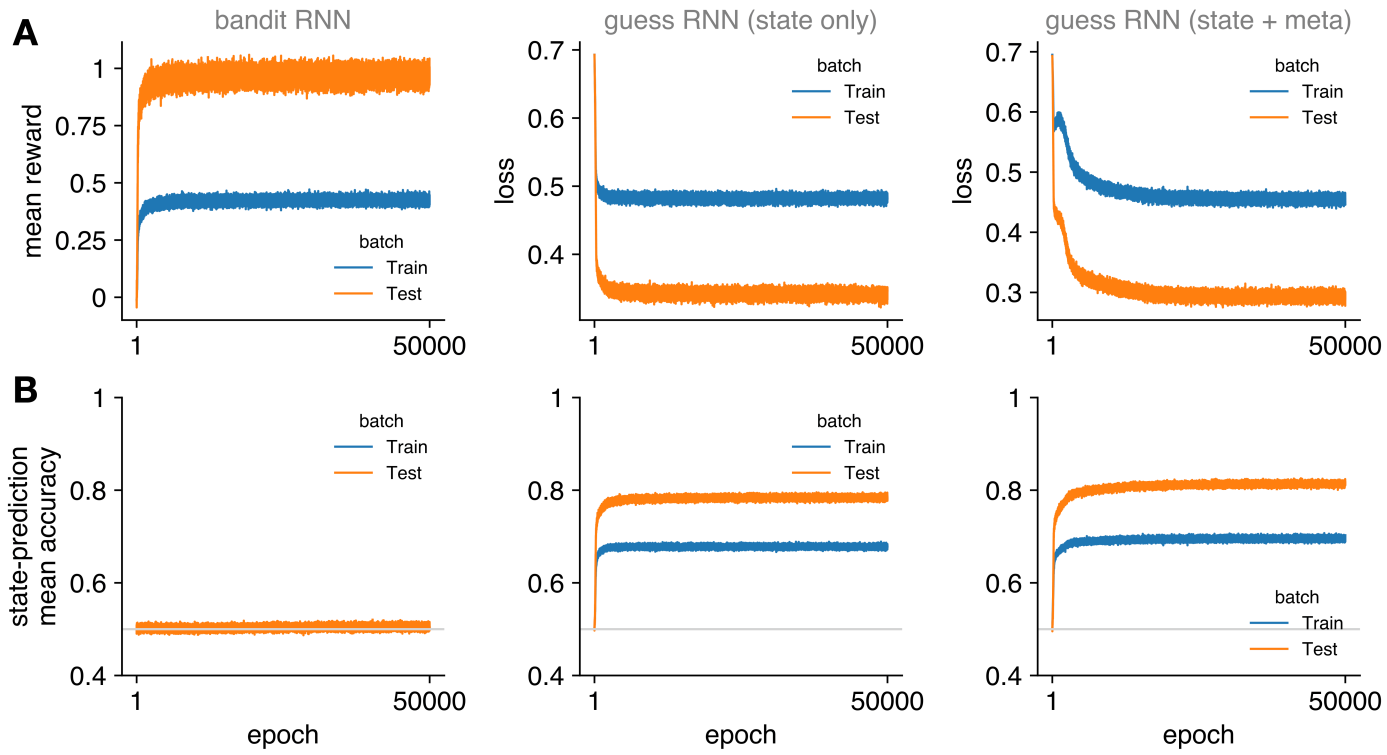
Supplementary Fig. S4. Sampling within the initial streaking phase. Trials are aligned to the first switch in each game. The x axis indexes position relative to the first switch, with 0 denoting the first choice of the alternative option (not plotted) and negative indices counting backward within the initial streak. For example, the choice sequence AAAAAB is aligned so that the B trial is at 0 and the preceding A trials appear at −1, −2, and so on. Plotted are mean signed evidence (A, B), reaction time (C), and p(repeat) (D) as a function of this relative index, shown separately for the two task conditions.



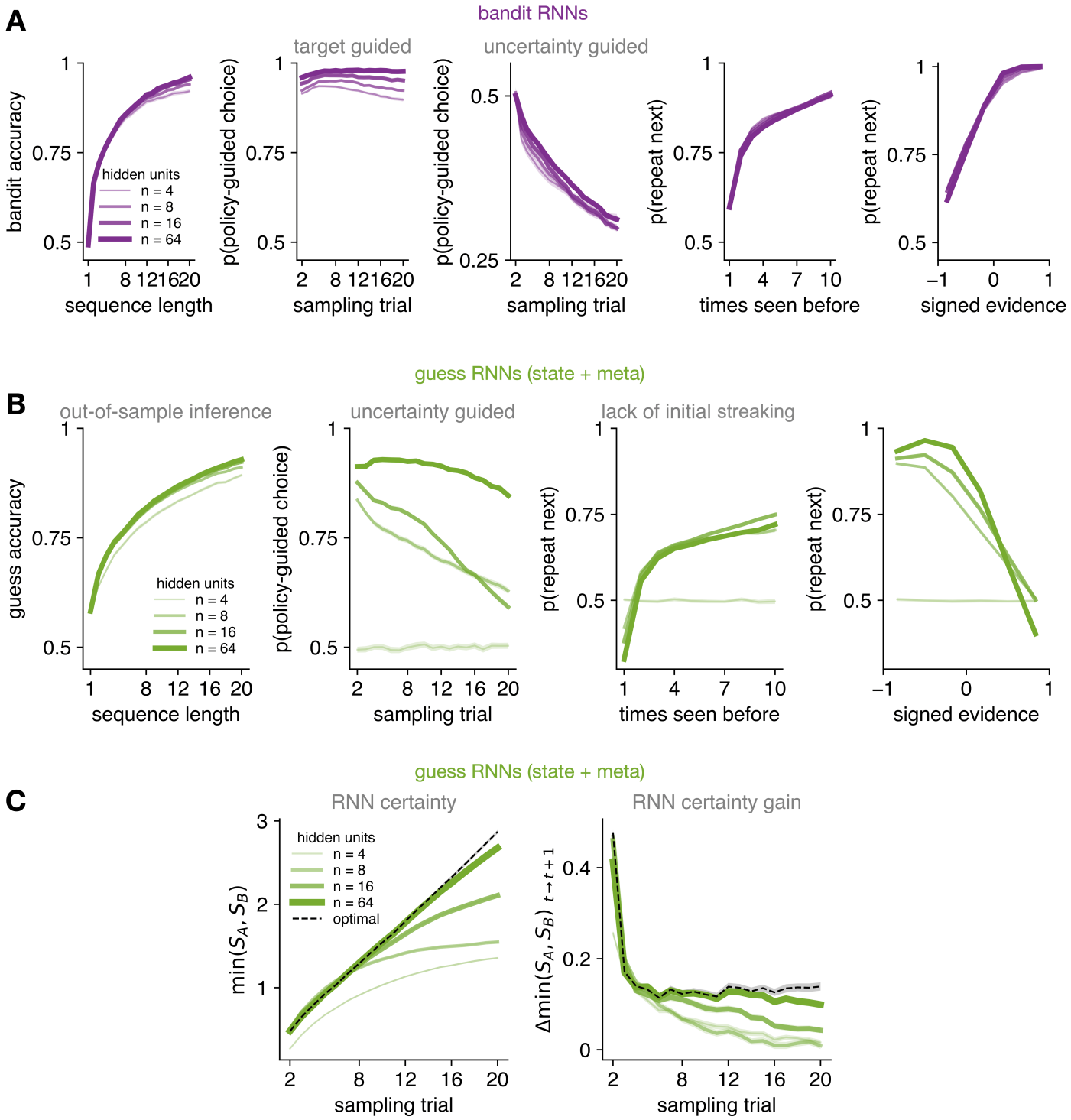
Supplementary Fig. S5. Random-effects Bayesian model comparison including variants with independent leak parameters for the chosen and unchosen options. The model space includes single-leak, dual-leak (δ chosen, δ unchosen), noise-only, and full leak + noise variants. Posterior model frequencies and family-wise comparisons are computed from individual ELBOs, separately for the bandit and guess conditions. Error bars indicate 95 percent confidence intervals.



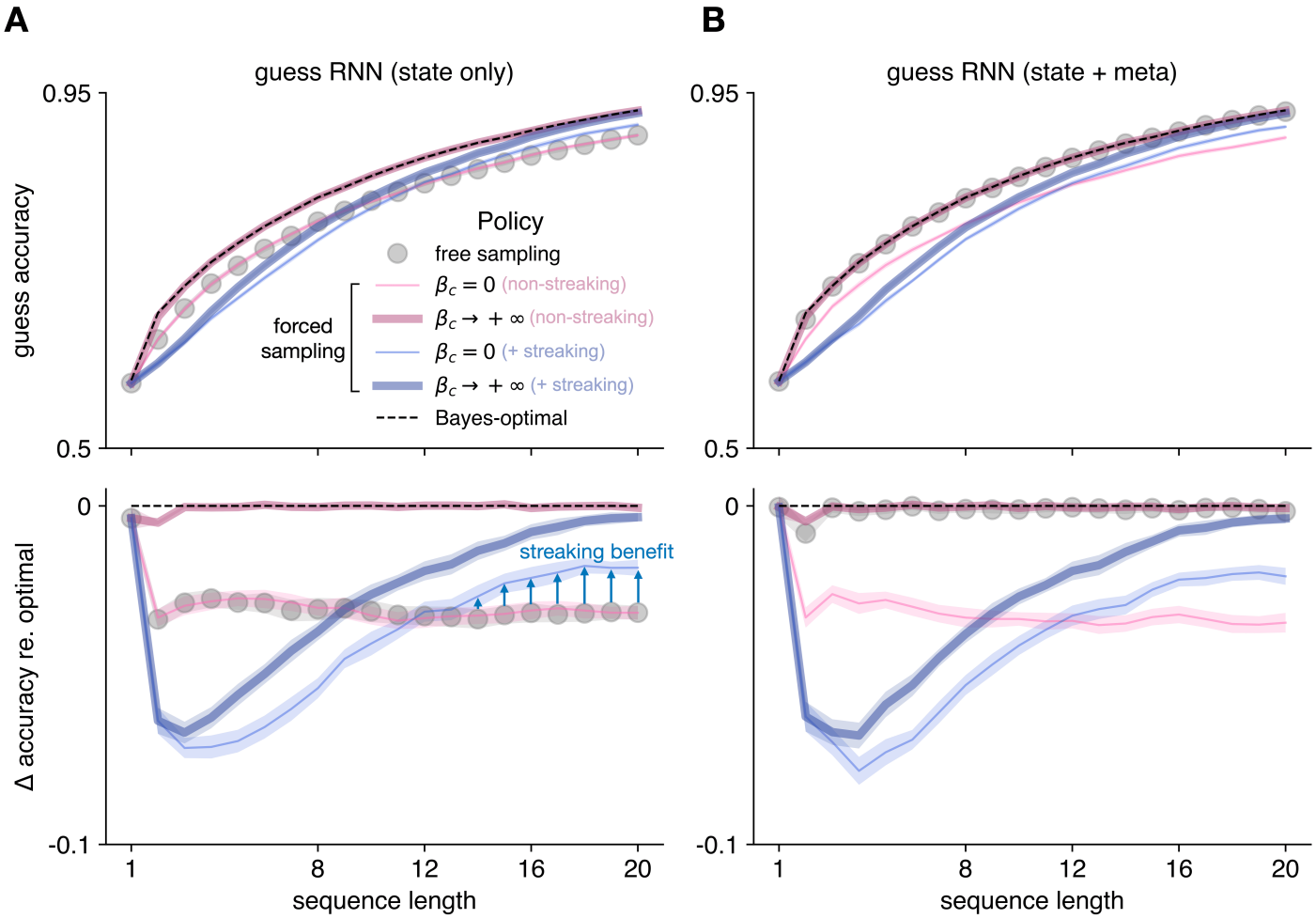
Supplementary Fig. S6. (A) Pairwise principal component plots of RNN hidden-layer activity. Game labels: o for orange bag, b for blue bag. Curves show trajectories across sampling trials; circle markers indicate sequence lengths of 8, 12, 16, and 20. (B) Single-trial correlations (top rows) and cross-validated variance explained (bottom rows) between each PC and the network's certainty lower bound, $\min(|S_A|, |S_B|)$, computed from the RNN output states.



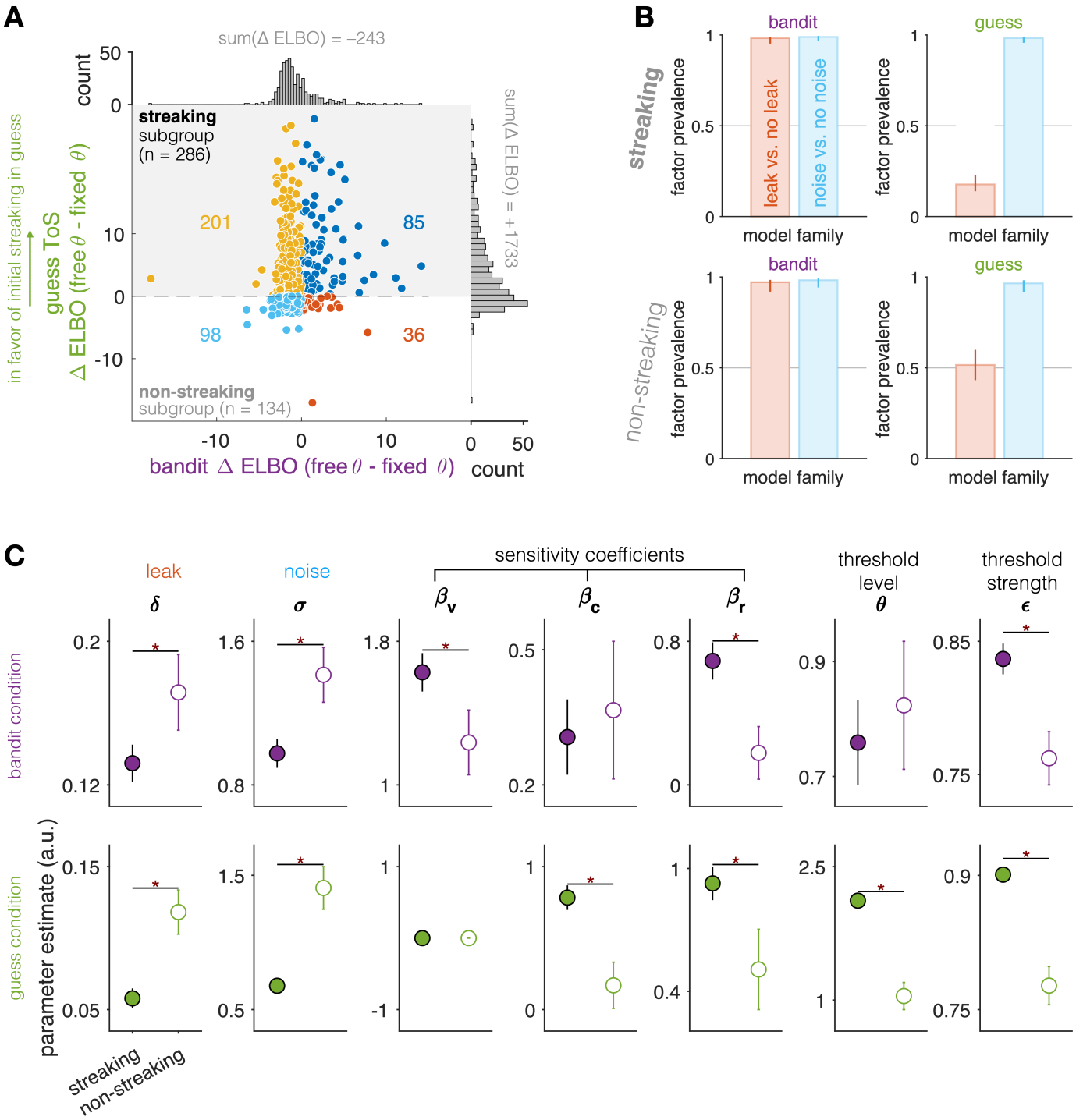
Supplementary Fig. S7. RNN convergence. (A) Training curves: mean reward per epoch for the bandit RNN and mean loss per epoch for the guess RNN. (B) State-prediction accuracy as a function of training epoch, evaluated on the training and test sets. Each training epoch comprised 50 sequences with variable lengths from 2 to 20 trials, sampled from a truncated geometric distribution with hazard rate 0.2. Sequence lengths were drawn until every length from 2 to 20 was represented at least once. Test data used fixed sequence lengths of 8, 12, 16, and 20 trials, matching the human experiment. Consequently, the training set contained, on average, a higher proportion of short sequences than the test set.



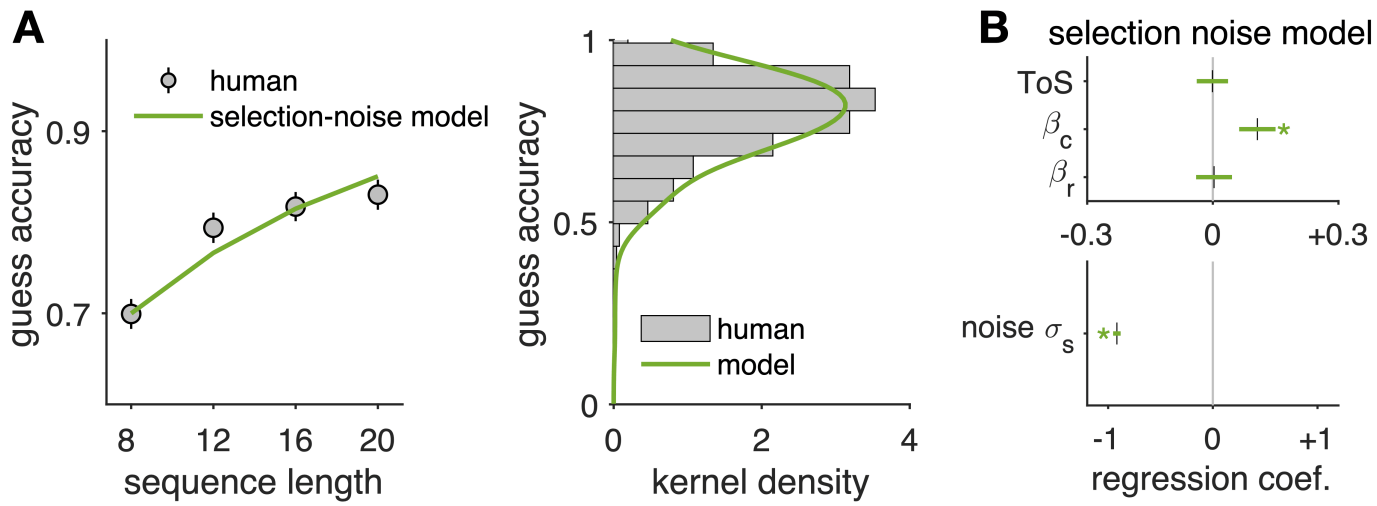
Supplementary Fig. S8. Effect of hidden-layer size on RNN performance. (A) Bandit RNNs are robust to hidden-layer size. (B) Guess RNNs with the state plus meta policy are sensitive to hidden-layer size for the emergence of uncertainty-guided sampling. With 4 units the network does not exhibit uncertainty-guided sampling, whereas with as few as 8 units it does. (C) Certainty signal, i.e., network's certainty lower bound, $\min(|S_A|, |S_B|)$, computed from the RNN output states. The 64-unit RNN approximates the Bayes-optimal benchmark.



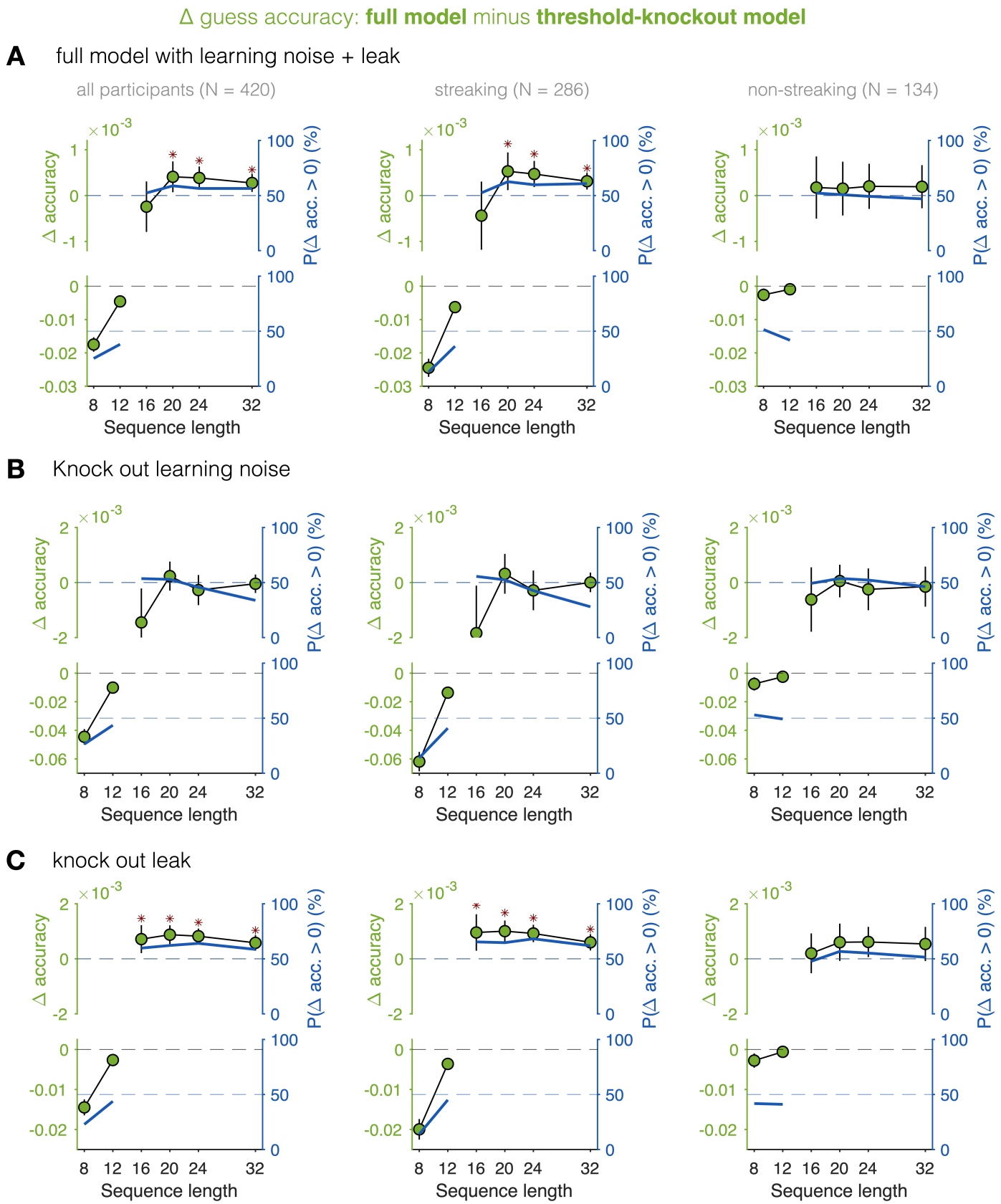
Supplementary Fig. S9. Guess accuracy from RNNs versus a Bayes-optimal agent (dashed). (A) Guess RNN trained for state prediction only. (B) Meta-policy RNN trained to predict states and to increase a certainty lower bound. Free sampling: choices generated by the RNN. Forced sampling: choices follow a predefined sequence from an external agent with $\theta = 2$ (streaking) or $\theta = 0$ (non-streaking), and either $\beta_c = 0$ (random) or $\beta_c \to \infty$ (perfect uncertainty guided). Bottom: accuracy shortfall relative to Bayes optimal.



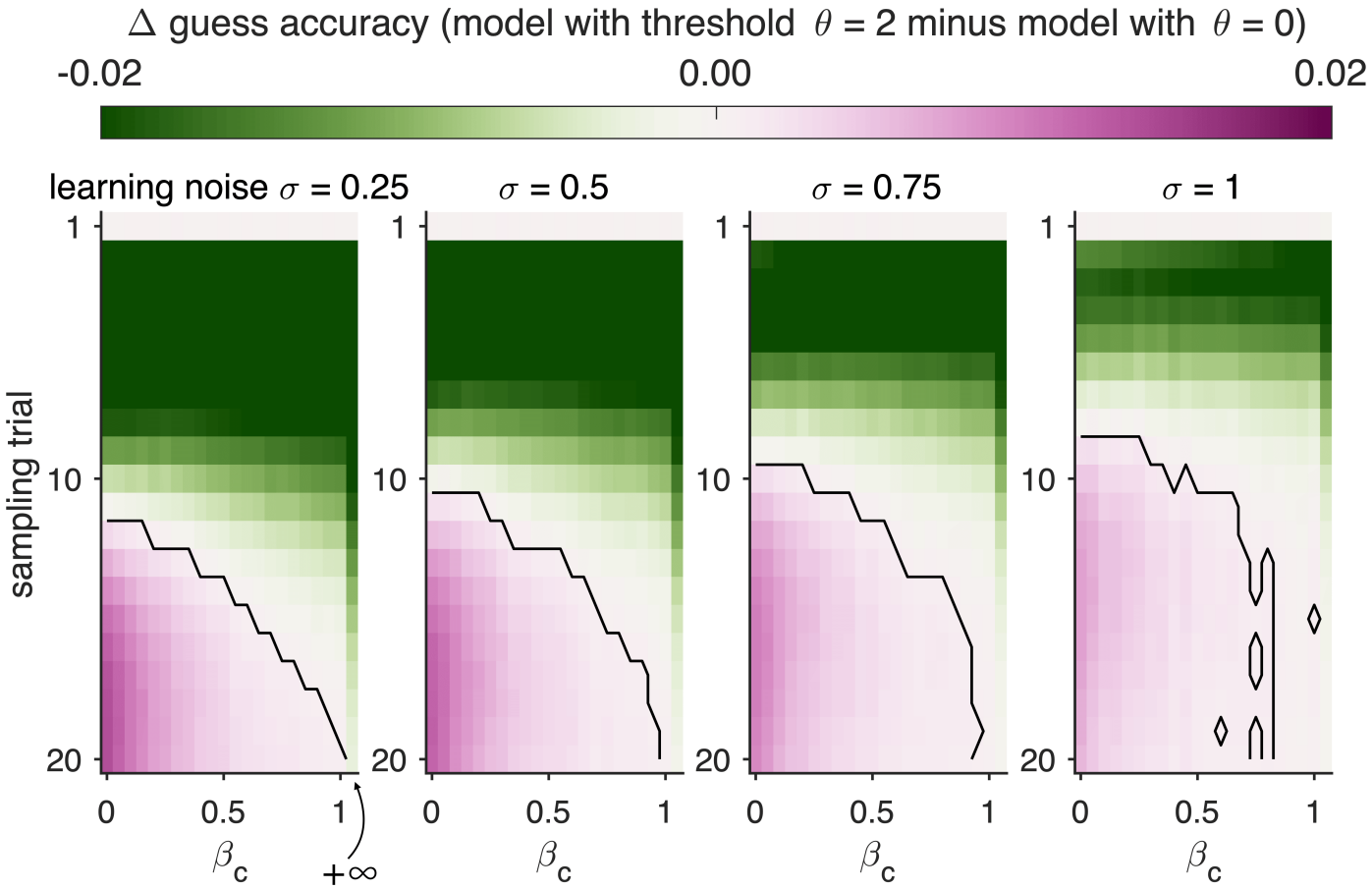
Supplementary Fig. S10. (A) Grouping 420 participants by model evidence ELBO in each condition. (B) Family-wise model factor analysis within streaking subgroup and non-streaking subgroup, respectively, for each task condition (bandit vs. guess). (C) Parameter contrasts between the two subgroups of participants in favor of an initial sampling phase (streaking) vs. without this initial sampling phase (non-streaking) in the guess condition.



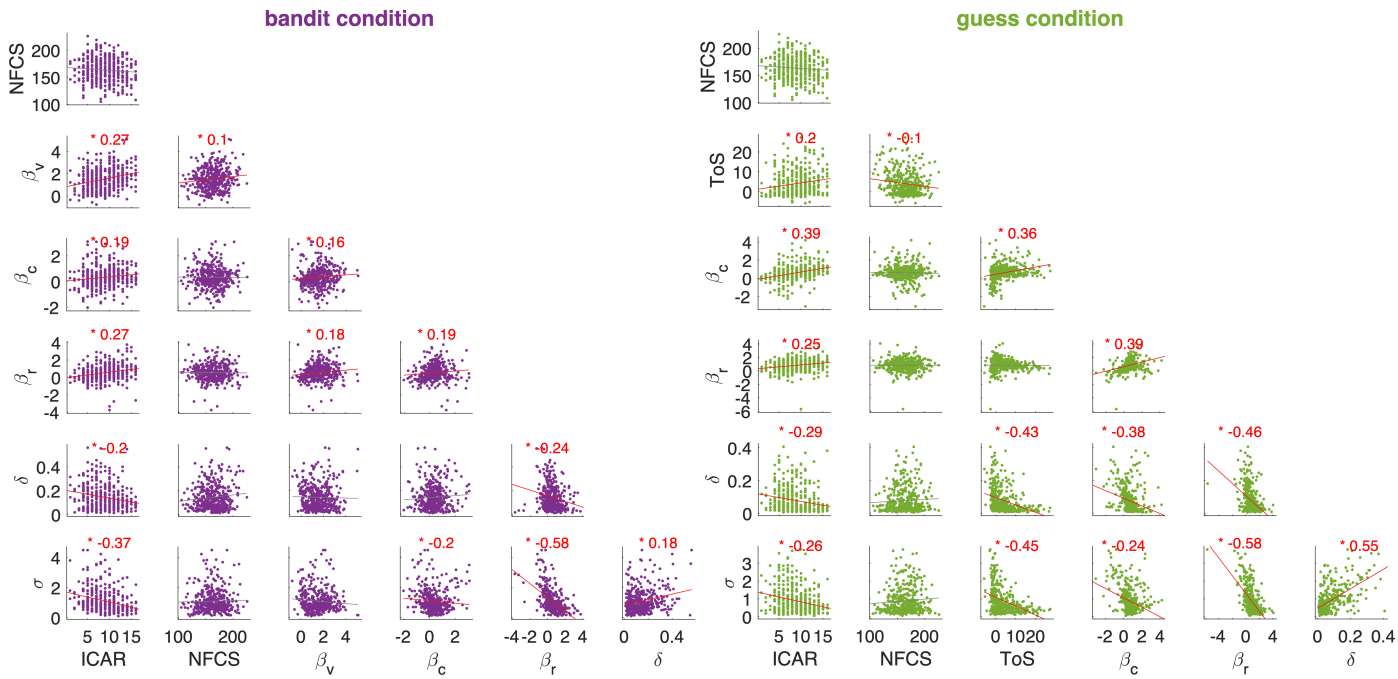
Supplementary Fig. S11. (A) Guess accuracy from a model without learning noise; instead, action selection noise perturbed posterior beliefs at the final probe of each game. Selection noise was drawn per participant from a gamma distribution with mean 3.5 and coefficient of variation 0.8, with parameters sampled independently to match the human distribution of mean and variance in guess accuracy. (B) In this model, only β_c positively predicted simulated accuracy, while ToS was not significant.



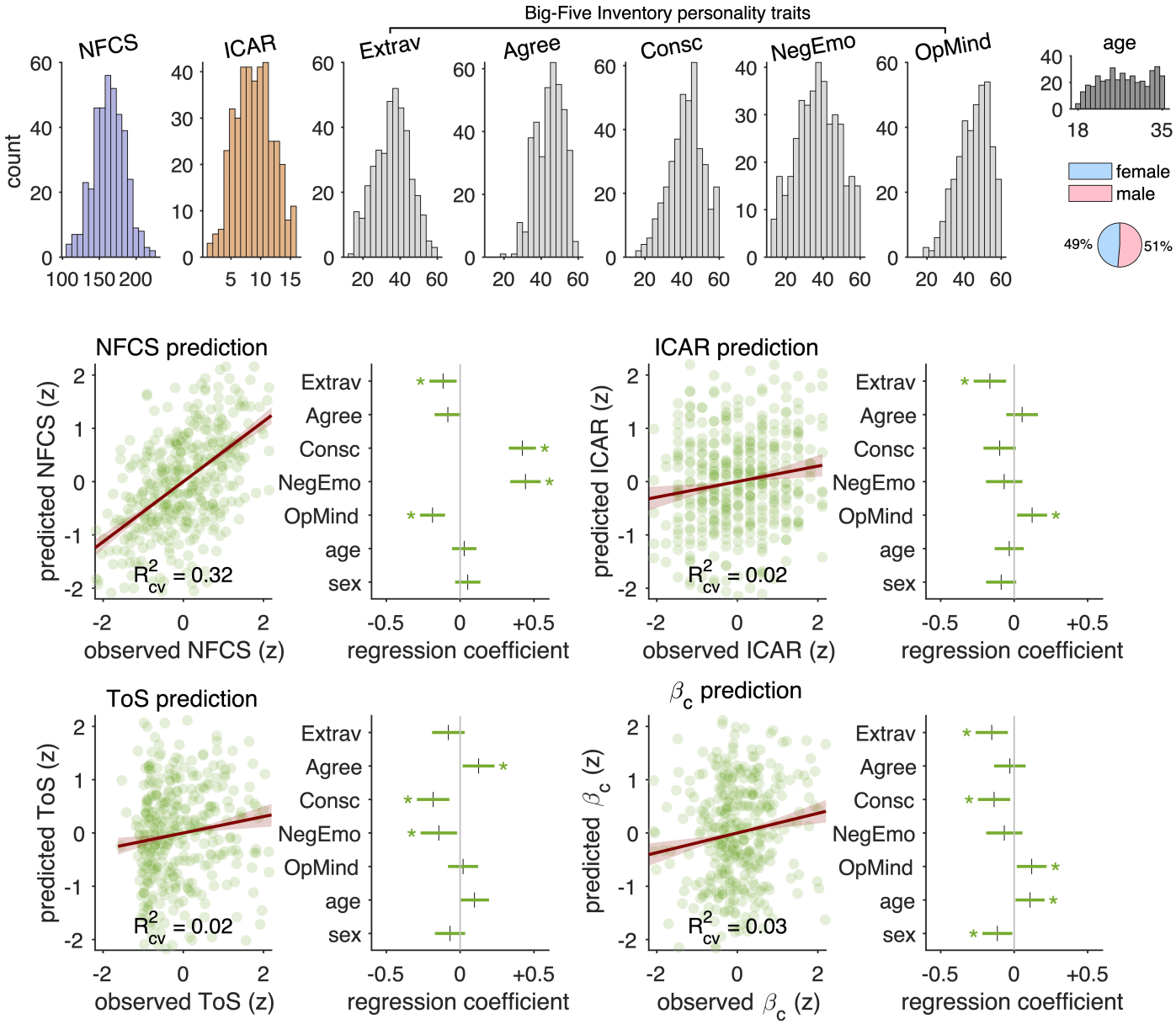
Supplementary Fig. S12. Benefit of streaking in the guess condition. (A) For each of 420 participants, guess accuracy was simulated (n = 1000 trajectories) using the best-fitting leaky-noisy model. A threshold-knockout model set $\theta = 0$ while keeping all other parameters identical. The streaking benefit is Δ accuracy = full model − threshold-knockout. (B) Same as A, with learning noise set to 0 in both models. (C) Same as A, with leak set to 0 in both models. Right y-axis (blue) shows the percentage of participants in the plotted cohort with Δ accuracy > 0. For longer sequences than in the experiment, stimulus sequences were duplicated to reach lengths 24 and 32 from 12 and 16. Dark red asterisks mark one-sided t tests against zero at $p < 0.005$.



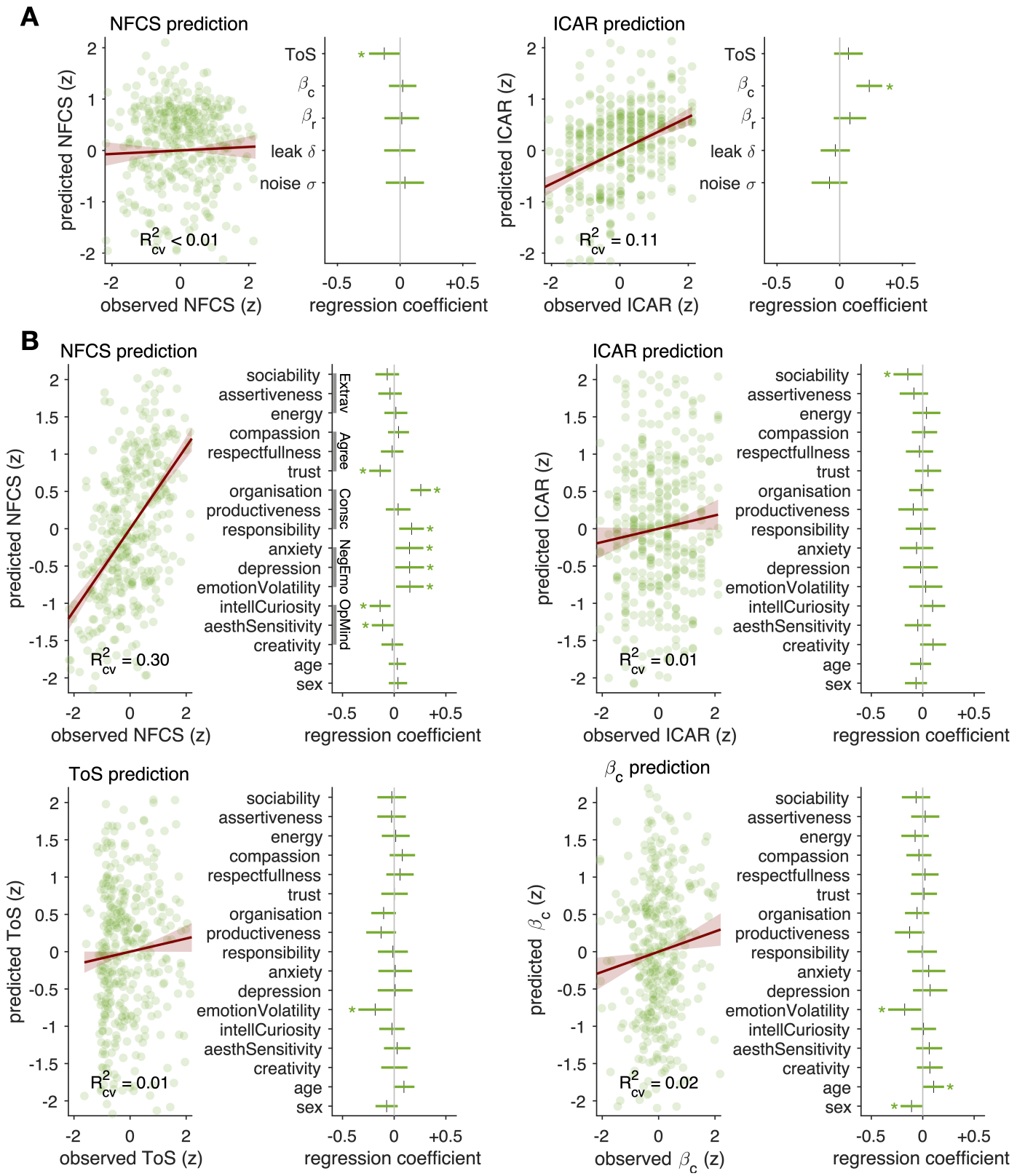
Supplementary Fig. S13. Increase in simulated accuracy from a non-streaking agent ($\theta = 0$) to a streaking agent ($\theta = 2$), with all other parameters fixed. Learning noise increases from left to right. The last column in each panel shows $\beta_c \rightarrow \infty$, corresponding to maximal sensitivity to uncertainty-guided sampling. The solid black outline marks clusters significant by a one-sided t test against zero at $p < 0.005$.



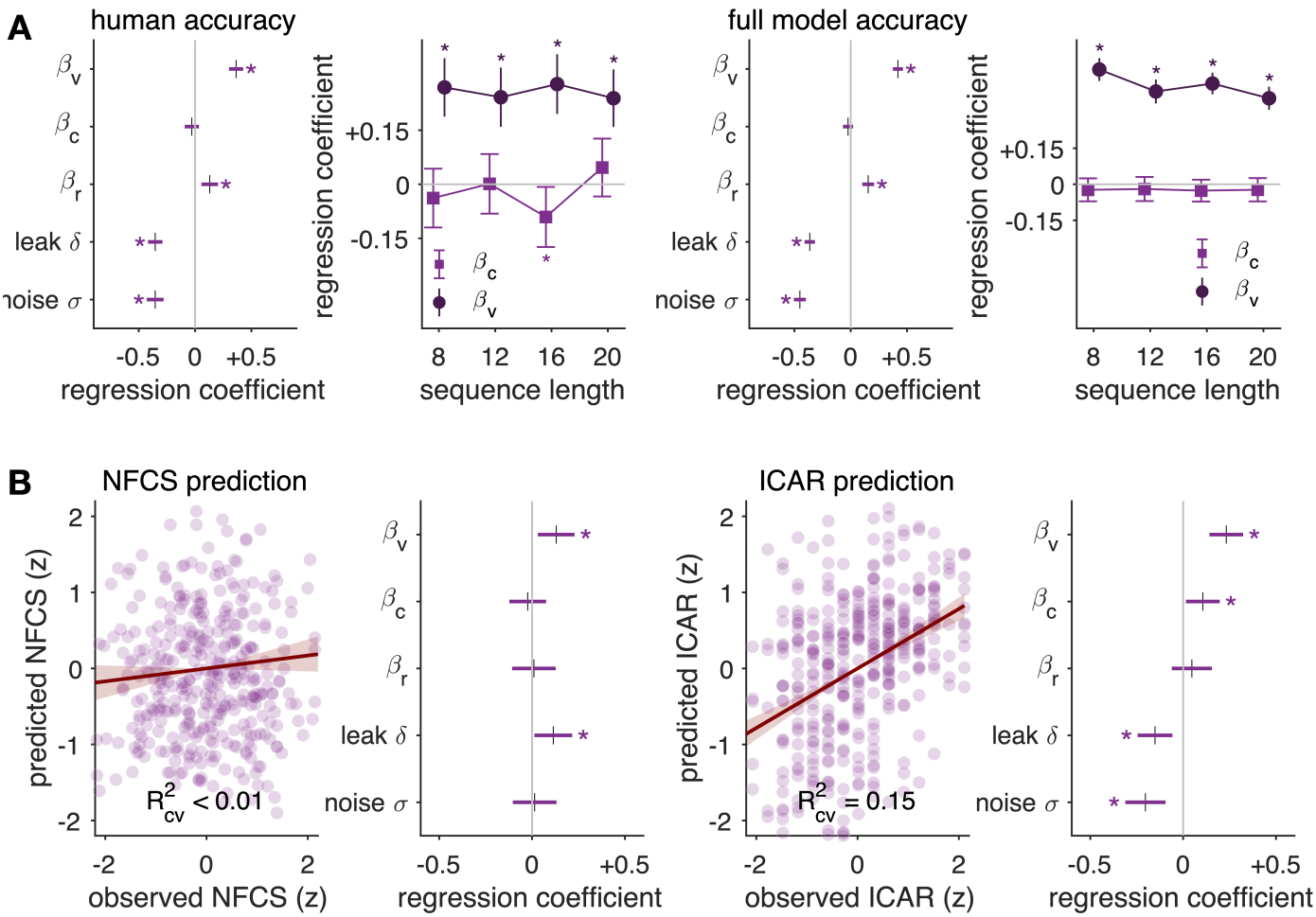
Supplementary Fig. S14. Inter-variable Spearman correlations among model parameters and NFCS/ICAR, in bandit and guess conditions.



Supplementary Fig. S15. **Personality trait taxonomy.** Leave-one-subject-out cross-validated prediction of NFCS, ICAR, β_c (uncertainty sensitivity), and ToS (streaking tendency) from Big-Five domain scores (Supplementary Fig. S16 for facets' scores). Prediction panels (left, for each variable): observed values versus cross-validated predictions; the line is the bootstrap mean linear slope with a 95% confidence band, and the cross-validated R^2 is reported at the bottom of each panel. Regression coefficient panels (right, for each variable): standardized regression coefficients by domain; the thin black vertical line marks the mean coefficient, the thick green horizontal bar shows the 95% confidence interval, and asterisks denote coefficients different from zero (two-sided one-sample t test).



Supplementary Fig. S16. (A) Leave-one-participant-out cross-validated prediction of NFCS and ICAR from model components in guess condition. (B) Leave-one-subject-out cross-validated prediction of NFCS, ICAR, ToS, and β_c from Big Five BFI-2 facet scores, with age and sex included as control covariates. Panels on the right side show standardized regression coefficients for all predictors. Panels on the left side show observed values versus cross-validated predictions; the line is the bootstrap mean linear slope with a 95% confidence band, and the cross-validated R^2 is reported in each panel. The thin black vertical line marks the mean coefficient, the thick green horizontal bar shows the 95% confidence interval, and asterisks denote coefficients different from zero (two-sided one-sample t-test).



Supplementary Fig. S17. (A) Predicting last-trial bandit accuracy in games with one target and one non-target shape from model parameters. (B) Leave-one-participant-out cross-validated prediction of NFCS and ICAR using parameters fit in the bandit condition.

Supplementary Table 1

Human accuracy ~

Model component	Mean coefficient	95% CI lower	95% CI higher	t value (df = 414)	p value
ToS	0.144	0.046	0.242	2.896	< 0.01*
β_c	0.211	0.121	0.300	4.623	< 0.0001*
β_r	0.080	-0.033	0.194	1.389	0.166
leak δ	-0.132	-0.232	-0.032	-2.597	< 0.01*
noise σ	-0.220	-0.345	-0.095	-3.454	< 0.001*

Model-predicted accuracy ~

Model component	Mean coefficient	95% CI lower	95% CI higher	t value (df = 414)	p value
ToS	0.063	0.029	0.097	3.635	< 0.001*
β_c	0.051	0.020	0.083	3.226	< 0.01*
β_r	0.008	-0.032	0.048	0.392	0.695
leak δ	-0.240	-0.275	-0.205	-13.533	< 0.0001*
noise σ	-0.742	-0.786	-0.698	-33.451	< 0.0001*

NFCS ~

Model component	Mean coefficient	95% CI lower	95% CI higher	t value (df = 385)	p value
ToS	-0.126	-0.248	-0.004	-2.031	0.042*
β_c	0.020	-0.089	0.130	0.368	0.713
β_r	0.013	-0.125	0.151	0.190	0.849
leak δ	-0.003	-0.127	0.122	-0.043	0.966
noise σ	0.038	-0.114	0.191	0.492	0.623

ICAR ~

Model component	Mean coefficient	95% CI lower	95% CI higher	t value (df = 385)	p value
ToS	0.070	-0.045	0.184	1.191	0.234
β_c	0.236	0.133	0.339	4.506	< 0.0001*
β_r	0.082	-0.048	0.212	1.241	0.215
leak δ	-0.035	-0.152	0.082	-0.592	0.554
noise σ	-0.081	-0.225	0.062	-1.116	0.265

Supplementary Table 2

NFCS to human accuracy, mediator = ToS, control variables = {age, sex}, (N = 391)

Path	Mean coefficient	95% CI lower	95% CI higher	t value	p value
a	-0.1294	-0.2411	-0.0177	-2.28	0.0233*
b	0.3267	0.2393	0.4141	7.35	< 0.0001*
direct c’	-0.0046	-0.0963	0.0871	-0.10	0.921
total	-0.0469	-0.1441	0.0504	-0.95	0.344
indirect a*b	-0.0423	-0.0802	-0.0062		0.0232*

ToS to human accuracy, mediator = NFCS, control variables = {age, sex}, (N = 391)

Path	Mean coefficient	95% CI lower	95% CI higher	t value	p value
a	-0.1258	-0.2354	-0.0161	-2.26	0.0247*
b	-0.0046	-0.0963	0.0871	-0.10	0.921
direct c’	0.3267	0.2393	0.4141	7.35	< 0.0001*
total	0.3273	0.2414	0.4132	7.45	< 0.0001*
indirect a*b	0.0006	-0.0124	0.0139		0.913

ICAR to human accuracy, mediator = β_c, control variables = {age, sex}, (N = 391)

Path	Mean coefficient	95% CI lower	95% CI higher	t value	p value
a	0.3135	0.2282	0.3989	7.22	< 0.0001*
b	0.2825	0.1878	0.3773	5.86	< 0.0001*
direct c’	0.3332	0.2398	0.4266	7.01	< 0.0001*
total	0.4217	0.3308	0.5127	9.12	< 0.0001*
indirect a*b	0.0886	0.0536	0.1294		< 0.0001*

β_c to human accuracy, mediator = ICAR, control variables = {age, sex}, (N = 391)

Path	Mean coefficient	95% CI lower	95% CI higher	t value	p value
a	0.3155	0.2148	0.4161	6.16	< 0.0001*
b	0.3332	0.2398	0.4266	7.01	< 0.0001*
direct c’	0.2825	0.1878	0.3773	5.86	< 0.0001*
total	0.3876	0.2850	0.4903	7.43	< 0.0001*
indirect a*b	0.1051	0.0673	0.1524		< 0.0001*

Supplementary Table 3

NFCS ~

Regressor	Mean coefficient	95% CI lower	95% CI higher	t value (df = 383)	p value
Extra	-0.116	-0.208	-0.024	-2.471	0.014*
Agree	-0.083	-0.173	0.007	-1.820	0.069
Consc	0.422	0.330	0.513	9.020	< 0.0001*
NegEmo	0.442	0.339	0.546	8.440	< 0.0001*
OpMind	-0.187	-0.272	-0.102	-4.323	< 0.0001*
age	0.028	-0.055	0.110	0.656	0.512
sex	0.051	-0.034	0.137	1.186	0.237

ICAR ~

Regressor	Mean coefficient	95% CI lower	95% CI higher	t value (df = 383)	p value
Extra	-0.165	-0.276	-0.055	-2.936	< 0.01*
Agree	0.054	-0.054	0.162	0.982	0.327
Consc	-0.098	-0.209	0.012	-1.749	0.081
NegEmo	-0.068	-0.192	0.056	-1.080	0.281
OpMind	0.122	0.020	0.224	2.347	0.019*
age	-0.033	-0.133	0.066	-0.663	0.508
sex	-0.089	-0.191	0.014	-1.698	0.090

ToS ~

Regressor	Mean coefficient	95% CI lower	95% CI higher	t value (df = 383)	p value
Extra	-0.080	-0.190	0.031	-1.420	0.156
Agree	0.125	0.017	0.233	2.276	0.0234*
Consc	-0.183	-0.293	-0.073	-3.259	< 0.01*
NegEmo	-0.144	-0.268	-0.021	-2.294	0.022*
OpMind	0.020	-0.082	0.122	0.384	0.701
age	0.097	-0.002	0.196	1.930	0.054
sex	-0.069	-0.171	0.034	-1.316	0.189

β_c ~

Regressor	Mean coefficient	95% CI lower	95% CI higher	t value (df = 383)	p value
Extra	-0.153	-0.263	-0.043	-2.731	< 0.01*
Agree	-0.029	-0.137	0.078	-0.537	0.592
Consc	-0.137	-0.246	-0.027	-2.446	0.015*
NegEmo	-0.067	-0.190	0.056	-1.075	0.283
OpMind	0.119	0.018	0.221	2.308	0.0215*
age	0.108	0.010	0.207	2.157	0.0316*
sex	-0.114	-0.216	-0.012	-2.207	0.0278*