Representational Constraints Shape Human Information-Seeking

Samuel J. Cheyette[1], Frederick L. Callaway[2], Neil R. Bramley[3], Jonathan D. Nelson[4], and Joshua B. Tenenbaum[1]

[1]Department of Brain & Cognitive Sciences, Massachusetts Institute of Technology
[2]Department of Psychology, New York University
[3]Department of Psychology, University of Edinburgh
[4]Department of Psychology, University of Surrey

Abstract

Human learners often seek information rationally: they prefer more informative over less informative data, and ask questions that optimally discriminate between competing hypotheses. But this picture holds only in simple, controlled settings: in more complex tasks, the computations needed for optimal information seeking become intractable and human inquiry empirically is often far from optimal. How do people effectively navigate the information landscapes the real world presents? We propose a model of information-seeking under representational constraints, which simplifies complex data used to make inferences and seeks information rationally with respect to its costs. Six behavioral experiments using two popular search games support this account: we find that people reason imprecisely and make queries in ways that are objectively uninformative under standard models, but are actually efficient given their limited representational capacity. Participants also increasingly prefer simpler, less informative queries as they gain experience in a task, suggesting rapid adaptation to their cognitive limits. Our findings challenge theories of human information-seeking based on an idealized notion of rationality and instead support a more nuanced and realistic view: people are limited in what information they can represent but have information-seeking strategies that are efficient given those limits.

Introduction

Philosophers, psychologists, and artificial intelligence researchers have long sought to ground human inquiry in rational terms [1–3]. Modern theories of information-seeking based on Bayesian principles of "Optimal Experimental Design" (OED) have had notable success explaining how people test hypotheses in simple settings, especially relative to other possible modes of inquiry like falsificationism [4, 5]. In certain circumstances, people ask questions and use problem-solving strategies that align with OED principles, devising queries that maximize expected information gain [3, 6–9]. However, other work has found that in other settings people seem to make queries that are informative about only a small number of hypotheses at a time [10–13], their questions are often simpler than information-maximizing questions [14, 15], and they frequently take more steps than necessary to solve a problem [5, 16], indicative of difficulty integrating evidence. An obvious challenge in formulating a unifying account of human active learning is reconciling these findings with the cases where human information-seeking aligns with OED principles. A hint about how this might be achieved is to note that the cases where people are found to be near-optimal are also generally the least complex tasks, involving only a small number of objects and hypotheses.

There are *prima facie* reasons why we might not expect people to do optimal information-seeking in complex settings, which become apparent when considering the computational resources inherent to optimal algorithms for active inference. In 1976, Donald Knuth wrote a paper entitled "The Computer as Master Mind," describing a computer program that optimally plays Mastermind [17], a game that involves figuring out a hidden code by making repeated guesses and using feedback to make deductive inferences. The algorithm works by maintaining a list of all possible solutions (1,296 of them), checking all the possible ways the game could continue after any possible query, and picking the one that minimizes the worst-case outcome. While people are more generally intelligent than any computer program (and were especially so in 1976), they find this game to be notoriously hard [16]: people take *on average* more guesses to reach a solution on a simplified version of Mastermind than Knuth's program does in the *worst case* on the standard version [18].

A key difference between humans and digital computers, which may help explain our shortcomings in this case, is that people's capacity for actively maintaining and manipulating information is quite small. A wide range of empirical evidence, dating back over a century, demonstrates that people have severely limited information-processing capacity in verbal memory [19], object tracking [20, 21], numerosity perception [22, 23], visual memory [24, 25], and semantic memory [26, 27], among many other domains. Attempts to measure the information content consistent with human performance across disparate behavioral tasks suggest that both visual and verbal short-term memory have a capacity of just a few or several bits [e.g. 19, 28–31]. For comparison, this is on the order of about 10^6 times less than the RAM of a 1970s-era computer [32] and about 10^9 times less than a modern laptop.

In this paper, we ask how a limited representational capacity affects people's ability to learn, reason, and seek information in the context of two logical strategy games. We consider two hypotheses jointly: first, that a limited capacity to represent information constrains how people interpret evidence; and second, that people strategically avoid information that exceeds their representational capacity in favor of simpler evidence. To test these hypotheses, we develop a formal theoretical framework for inference and information-seeking under representational constraints. Assuming zero representational cost, the model is equivalent to Bayesian inference and utility-maximizing (similar to OED); however, assuming non-zero cost, the model represents the implications of available evidence only approximately and seeks information that respects those imperfect approximations. We then compare human performance in six experiments involving two logical strategy games to this model, as well as to models that randomly query, information-maximize with no constraints, and information-maximize but lose information over time. We find that people make queries that provide simpler answers than would be optimal under a global information-maximizing perspective—however, their queries are close to information-maximizing when accounting for cognitive constraints.

Framework

To help motivate our modeling approach, consider the example of a doctor faced with diagnosing a patient who presents with a variety of symptoms and a set of test results. Each piece of evidence (fever, rash, elevated white blood cell count) is informative about the probability of different diagnoses. Ideally, the doctor would consider the full implications of each test or symptom — how each piece of information changes the probability of all possible diagnoses. In practice, however, the doctor may use simplified interpretations of each symptom or test, such as thinking "elevated white blood cells are consistent with a possible infection," even though there are certain infections that do not result in elevated white blood cell counts and there are other causes of elevated white blood cell counts. These simplified interpretations of evidence may contribute to uncertainty about a diagnosis even when it is completely determined by perfect reasoning about all available information. Our model aims to capture this kind of imprecision in how people use information to shape their beliefs, as well as how people may adapt their information-seeking behavior to account for that imprecision.

Approximate representations of evidence

Suppose we have a set of mutually exclusive hypotheses, $H = \{h_1 \dots h_n\}$, for which we have some prior distribution $P(H)$. The rational way of updating our beliefs upon receiving information E is to use Bayes' rule, i.e. to compute the posterior $P(H \mid E) = \frac{P(E|H)P(H)}{P(E)}$. However, this computation can be quite costly, and past work has explored different mechanisms by which people may approximate the Bayesian ideal [33–36]. Here, we focus on one specific challenge for Bayesian updating: understanding the implications of the data itself. Formally, the implications of data are captured by the *likelihood function*, $P(E \mid H)$, which

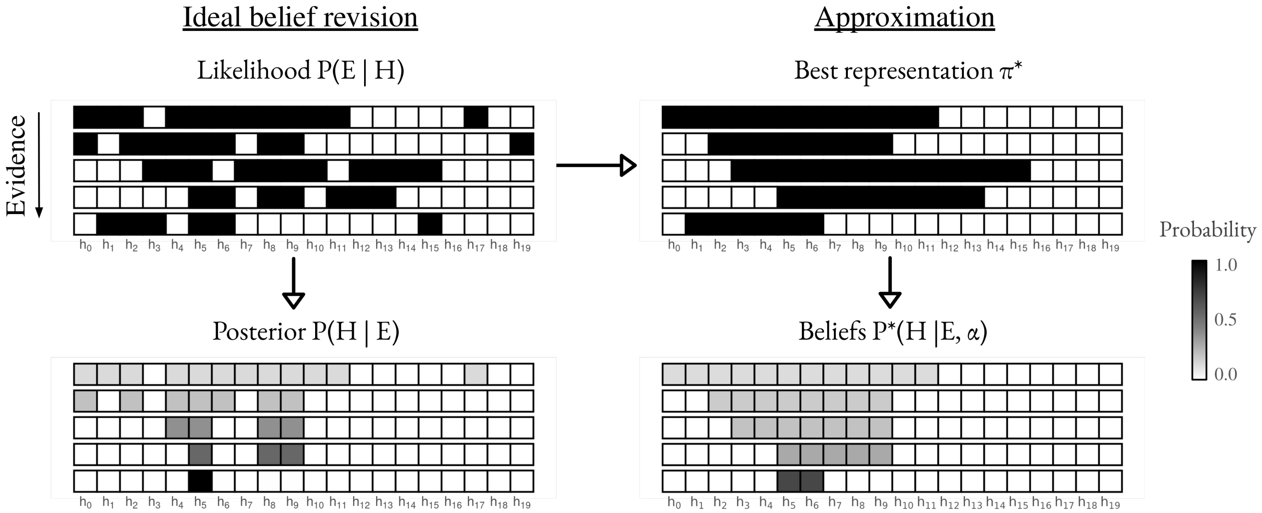


Figure 1: A conceptual illustration of the model with a binary likelihood. In each panel, the columns correspond to twenty mutually exclusive hypotheses ($h_0 ... h_{19}$) and the rows correspond to distinct pieces of evidence, presented sequentially from top to bottom. The top left panel shows the true likelihood of hypotheses given the evidence, with a black cell indicating that the evidence is consistent with the given hypothesis (note that the likelihood is implicitly conditioned on a query). The bottom-left panel illustrates perfect integration of that evidence over time, with more likely hypotheses having a darker shade. The top right panel illustrates one possible imperfect representations of the likelihood—the best contiguous range of 1's (black) in terms of overlap with the likelihood. The bottom right panel shows the corresponding posterior beliefs; note that the resulting distributions do not contain all available information.

specifies the probability of data under different hypotheses. Crucially, however, people must first find and represent this function before they can integrate it into their beliefs. To formalize this, we assume that the likelihood associated with evidence, E, is represented by a *program*, π_E, which takes a hypothesis as input and returns an (approximate) likelihood; that is, $\pi_E(H) \approx P(E \mid H)$. We assume that there is a cost involved in finding and representing a program proportional to its complexity, which we denote $K(\pi)$.

Given a sufficiently expressive family of programs Π, there will always be some program $\pi \in \Pi$ that exactly matches any given likelihood function, i.e., where $P(E \mid H) = \pi(H)$. However, this program may be very complex (long), incurring high cost. To avoid this cost, one could use a simpler program, but it will only imperfectly capture the true implications of the data. We quantify this discrepancy with a distance metric $D\left(P(E \mid \cdot), \pi(\cdot)\right)$. To capture the tradeoff between the program's accuracy and its cognitive cost, we define the utility of representing the likelihood of evidence E with program π as

$$U(\pi, E; \alpha) = -D\left(P(E \mid \cdot), \pi(\cdot)\right) - \alpha K(\pi), \tag{1}$$

where α is a scaling parameter that governs the relative weight of the two types of cost (inaccuracy and complexity). The bounded-optimal solution is to pick the highest utility program, $\pi^* = \arg\max_{\pi \in \Pi} U(\pi, E; \alpha)$. However, we model people's approximate optimization with a softmax, $P(\pi \mid E, \alpha, \beta) \propto \exp(\beta U(\pi, E; \alpha))$.

In order to perform inference, we modify Bayes' rule to account for approximations to the likelihood function. We will denote the approximate posterior distribution as $P^*(h \mid E, \alpha, \beta)$ to discriminate it from the true posterior probability $P(H \mid E)$. This is given by

$$P^*(h \mid E, \alpha, \beta) \propto P(h) \left(\sum_{\pi \in \Pi} P(\pi \mid E, \alpha, \beta)\, \pi(h) \right). \tag{2}$$

In other words, given prior belief $P(h)$ that a hypothesis is correct, the approximate posterior distribution after receiving new information E is the product of the prior and the program-approximated likelihood, where the programs are weighted by their probabilities.[1] Note that as α gets small, $P(\pi \mid E, \alpha, \beta)$ will increasingly favor programs that exactly match the true likelihood. Thus, as long as the set of possible programs Π is sufficiently expressive, $P^*(H \mid E, \alpha, \beta)$ will converge to the true posterior $P(H \mid E)$. Figure 1 provides a conceptual illustration of this idea.

Active learning

Models of active learning typically assume that people ask questions, make queries, and perform experiments that aim to maximize their expected information gain [e.g. 5, 6, 11, 37–40] — often called the 'Optimal Experimental Design' theory, since it assumes that people act like scientists who perform tests to best discriminate between hypotheses. A standard model, which we will adopt here, assumes that a query is valuable to the extent that the expected KL-divergence between the prior and posterior (the mutual information) is high. Intuitively, a good query is one whose outcome is likely to produce a large change in your beliefs. However, if people use simple approximations to interpret the implications of their observations, the actual change in belief will be different from the ideal Bayesian update. This motivates our key question: Are people sensitive to this difference, and do they adjust their information-seeking behavior accordingly? If people are attuned to their own representational limits, they should seek less complex information than if they were capable of unbounded representation. Specifically, a boundedly rational agent would aim to maximize the expected information gain with respect to their imperfect belief-updating,

$$U(\text{query}) \propto \sum_{h \in H} P(h) \sum_{o \in O} P(o \mid \text{query}, h) D_{KL}\left[P^*(H \mid o, \alpha, \beta) \parallel P(H)\right], \tag{3}$$

[1] By marginalizing over programs, we are making a mean-field approximation with respect to the stochastic choice of program.

where H are hypotheses, O is the set of outcomes one could observe from a given query, and $P^*(H \mid o, \alpha, \beta)$ is the distribution resulting from approximate belief updating.

Gameplay with limited representation

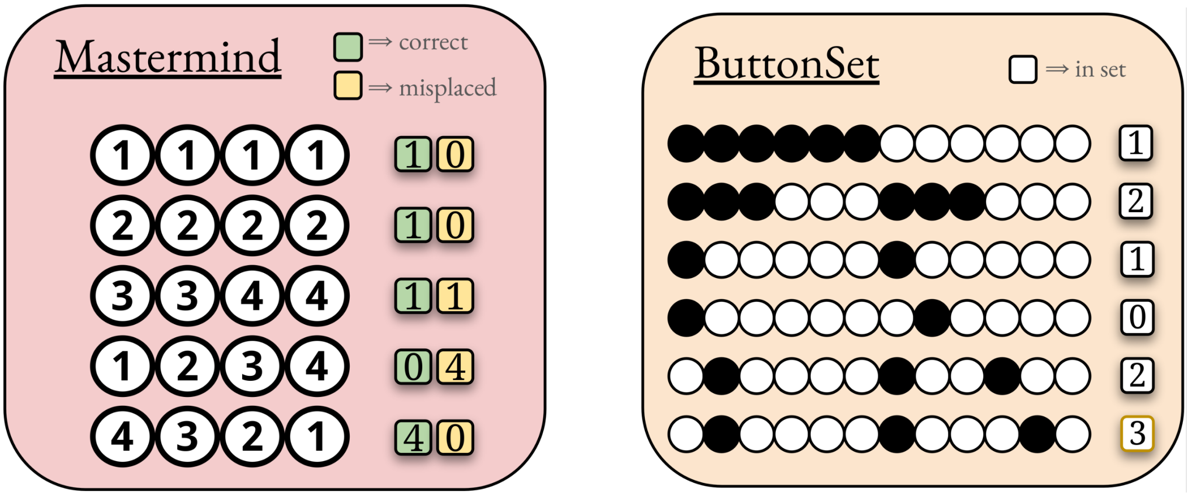
We now describe the results of applying this model to two games, which we also use for subsequent experiments with adult participants: a variant of the classic game Mastermind and a novel game we call ButtonSet. Both games have a similar structure that involves making repeated queries and using feedback from those queries to help infer an unknown code (Mastermind) or set (ButtonSet). The structure of these games is of theoretical interest because it involves several essential elements of real-world learning [41]: 1) there is an unknown state of the world that the agent wants to identify; 2) the agent can perform tests or make queries; 3) there are rules to the world that determine the relation between test and outcome; 4) the agent must use observations resulting from the tests or queries, along with an understanding of the rules, to make inferences that aid in narrowing the space of possible world states.

Both games also require non-trivial inferences regarding the implications of information received from making queries, making them useful tests of the model. To make the model concrete, we define the space of programs, Π, using a grammar over first-order logical expressions. Each expression corresponds to a logical statement about properties of the unknown code or set, for example, "there is exactly one **1** in the code." See Supplementary Tables 1 & 2 for full specifications of the grammar for each game. Following the model above, we assume that people may approximate the likelihood (the implications of feedback in the games) with simpler expressions. We define the complexity of a program, π, as the amount of information needed to represent π under grammar G under an optimal coding scheme; that is, $K(\pi) = -\log(P_G(\pi))$.

Mastermind The goal of Mastermind is to guess a code consisting of four colors or digits (here we use digits)[2] by repeatedly making queries and receiving feedback. Each time a player makes a guess they receive two pieces of feedback: how many digits of their guess are in the code at the correct position and how many are in the code but in the wrong position. So, if a player guesses 1233 but the true code is 1344, they will learn that one digit of their query is in the code at the right position (in this case the 1) and one digit is in the code but at the wrong place (in this case the 3). They are not told which digits were correct or partially correct, only how many of each.

Table 1 gives an example of logical expressions that approximate the meaning of feedback in a game of Mastermind (using 3-digit, 3-slot Mastermind for simplicity). For the first guess, 111, there is a short expression that exactly captures the meaning of the feedback, that one element is in the correct position and zero are misplaced: $\exists! x \in C\ (x = 1)$. However, for the

[2] In the original game of Mastermind, there are six allowable digits (or colors) in each of four slots. We only allowed four digits (1-4) in the four slots here.



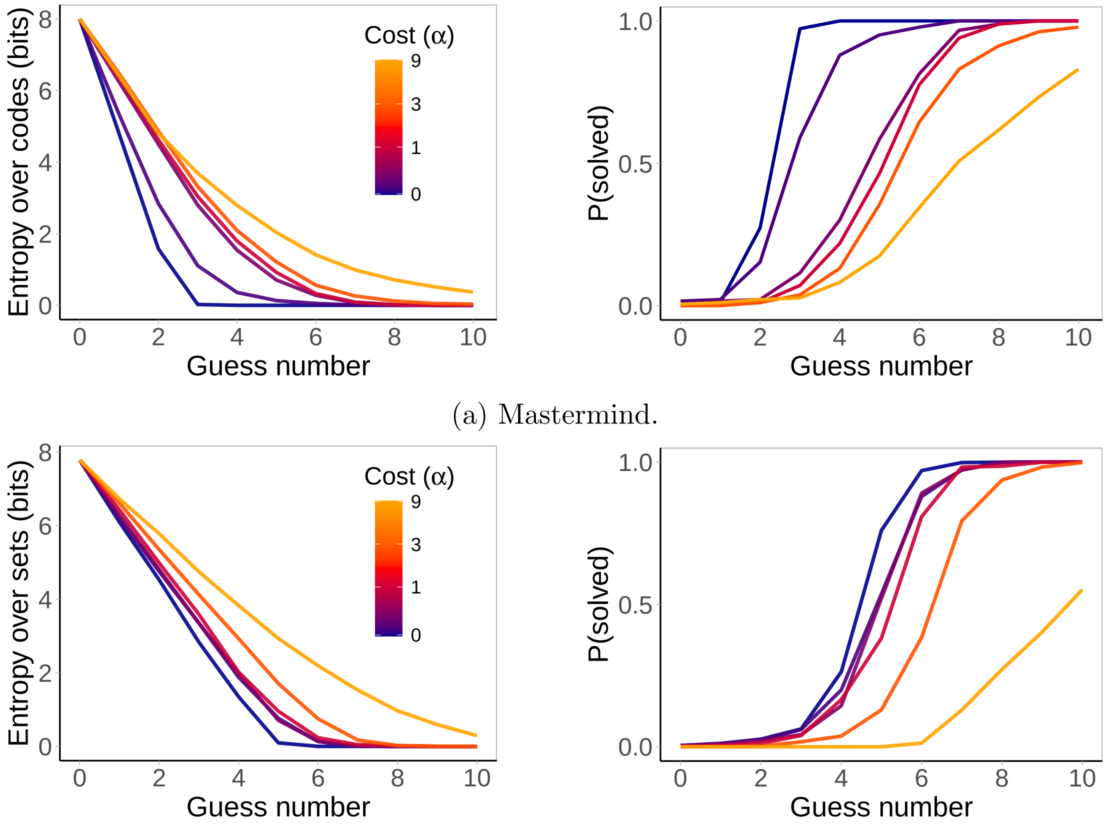
(a) (b)

Figure 2: Games of (a) Mastermind and (b) ButtonSet played to completion in our experimental setup. The goal of Mastermind is to guess a 4-digit code; each time a player makes a guess, they learn how many digits are correct (green) and misplaced (yellow). The goal of ButtonSet is to guess the the 3 true buttons out of a panel of 12 buttons. On each turn, the player can select a subset of up to 6 buttons and the feedback tells them how many of the true set of buttons is in the guess.

Guess	C	M	Approximation	Meaning
111	1	0	$\exists! x \in C \ (x = 1)$	There is exactly one 1 in the code.
321	0	2	$\neg\,((C_2 = 2) \vee (C_3 = 1))$	2 is not in position two and 1 is not in position three.
213	2	0	$(C_2 = 1) \vee (C_1 = 2)$	1 is in position two or 2 is in position one or both.
212	3	0		(Game Over)

Table 1: Reasoning in a 3-digit, 3-slot Mastermind game using approximate logical representations. The first column shows a guess, the second and third columns show the feedback (number correct (C) & number misplaced (M), respectively), the fourth column shows a logical expression partially capturing the meaning of the feedback, and the final column gives a textual description of the logic.

second and third queries, the shortest exact logical expression capturing the feedback are much longer, essentially just disjunctive statements about which codes are consistent. But there are much shorter approximate statements that *nearly* match the semantics of the feedback. For the third query, 213, the feedback of 2 correct and 0 misplaced can be approximated by $((C_2 = 1) \vee (C_3 = 2))$, which means that either 1 is in position two or 2 is in position three or both.



(a) Mastermind.

(b) ButtonSet.

Figure 3: Model simulation results for (a) Mastermind and (b) ButtonSet. The left panels show the model's average uncertainty in bits about the code/set (y-axis) after each guess (x-axis). The right panels show the probability of having solved the game (y-axis) after each guess (x-axis). Each line (color) represents a different value of the cost parameter α. As α gets closer to 0, the model makes higher-information queries and solves the game more quickly.

ButtonSet In addition to Mastermind, we created a novel game which is similar in structure to Mastermind but simpler in multiple ways. The game involves a player attempting to discover a hidden set of buttons from a spatially contiguous buttons, in this case consisting of 3 "true" buttons out of a panel of 12 buttons. On each turn, the player can select a subset of up to 6 buttons. After making a selection, the player receives feedback indicating how many of the selected buttons are part of the true code. The game continues until the player correctly identifies the exact set of 3 true buttons. Unlike Mastermind, the feedback in this game only specifies the number of correct selections without distinguishing their positions, making it both less informative but likely easier to reason about. It is also quite similar to a causal learning task in which the goal is to figure out which among a set of switches is working or broken by flipping subsets of them to see if a light turns on [42], the only difference being that in ButtonSet the feedback isn't binary (0/1) but rather a number between 0 and 3.

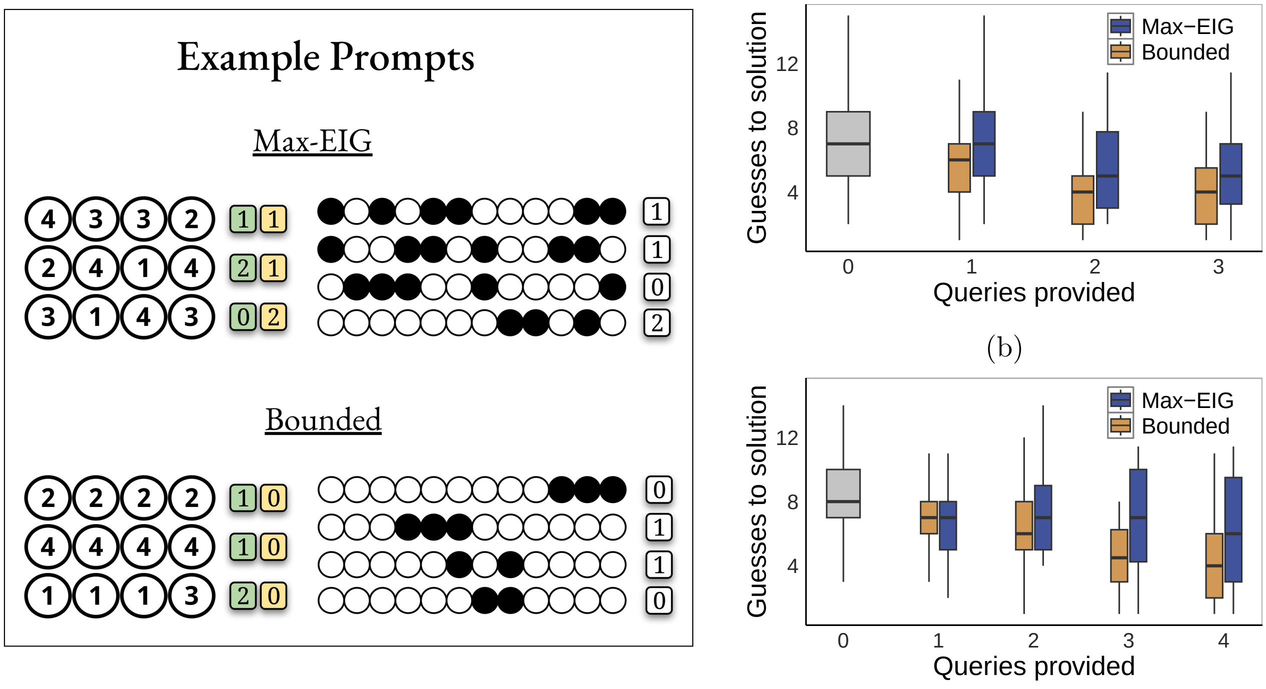
Model predictions Figure 3a-b shows the results of model simulations in both games. As α increases, the model becomes more sensitive to the complexity of the approximating functions it uses, leading to slower reductions in entropy and longer paths to a solution. This is both because when representational costs are high, the model uses less precise approximating functions, so draws less precise inferences; and also because it makes queries that return less complex feedback when α increases since it is optimizing around its own information-gain rather than true informativeness.

Experiment 1: Faster learning from simpler data

We first set out to test a key prediction of our model: that people will in certain cases learn *less* from more informative data relative to simpler data. We asked participants to complete games starting with an initial set of queries and results (0-3 queries for Mastermind and 0-4 for ButtonSet). These queries were always generated by maximizing expected information gain (EIG) given the previous evidence. However, in the "Unbounded EIG" condition, EIG was computed assuming perfect unbounded inference; in the "Bounded EIG" condition, EIG was instead computed under our bounded representation model. If participants are primarily limited in how they represent information, they will perform better in the Bounded EIG condition; if they are instead primarily limited by their ability to find informative queries, they will perform better in the Unbounded EIG condition. Importantly, this latter prediction holds even if people's inferences are subject to unstructured noise (the "leaky posterior model"), since effective EIG for this model is simply a fixed proportion of the true EIG. Figure 4 shows examples of prompts in both games in each condition (Max-EIG vs. Bounded). We recruited 50 participants for each game, each of whom played 5 rounds in each condition. After excluding those who solved fewer than half the games, we retained 44 participants (Mastermind) and 45 participants (ButtonSet).

Results

In both Mastermind and ButtonSet, the initial entropy predicted the number of guesses required to solve a game, meaning that participants made use of the information from the provided queries (MM: $B = 0.62$, $t(283) = 4.94$, $p < .001$; BS: $B = 0.77$, $t(387) = 8.22$, $p < .001$). However, participants solved games significantly faster when starting from simpler queries compared to Max-EIG queries. Games in which simple prompts were provided took fewer additional steps to solve compared to those with Max-EIG queries (MM: $B = -2.29$, $t(283) = -5.89$, $p < .001$; BS: $B = -1.19$, $t(387) = -4.04$, $p < .001$). This indicates that participants were less effective at using the information provided by the Max-EIG queries in both games, even though those queries were more objectively informative. In both games, participants took significantly longer to finish the game even when entropy was zero in the Max-EIG condition, indicating they struggled to fully capitalize on the information (see Figure 4b and



(a)

(b)

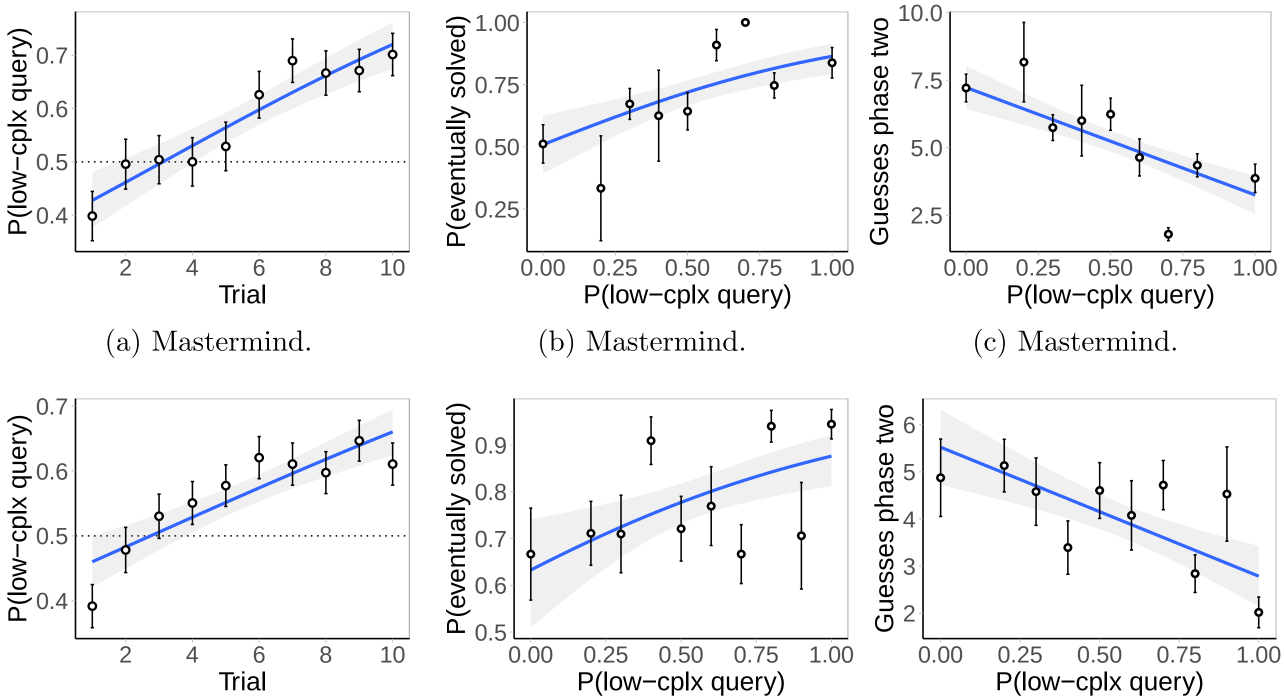
(c)

Figure 4: (a) Examples of model-based prompts that initiated games in Experiment 1. On the top are prompts from the Max-EIG model; on the bottom are prompts from the Bounded model, which penalizes representational complexity/. (b) Results of Experiment 1 for Mastermind; (c) Results of Experiment 1 for ButtonSet. Both panels show the number of guesses participants took to solve the game (y-axis) as a function of the number of initial queries provided (x-axis), split by how those queries were generated (color). The gray bar shows cases where no initial queries were provided; the blue bars show cases where the provided queries were maximally informative; and the orange bars show cases where provided queries traded off between informativeness and simplicity.

4c).

Experiment 2: Learning to pick simpler queries

Experiment 1 showed that participants did not effectively process feedback resulting from information-maximizing queries relative to simpler, less informative queries ones in either game. In this study, we aimed to test whether participants would actively *choose* less complex and less informative queries over maximally informative queries when presented with these options in a forced-choice paradigm. Participants played a modified version of both Mastermind and ButtonSet, each game consisting of two phases. In phase one, participants were repeatedly given two options to select between as a guess. This continued until the true code or set was perfectly determined. In phase two, participants tried to guess the code or set



(a) Mastermind. (b) Mastermind. (c) Mastermind.

(d) ButtonSet. (e) ButtonSet. (f) ButtonSet.

Figure 5: Results from Experiment 2 (Mastermind: a-c, ButtonSet: d-f). (a,c) The probability that participants picked the low-complexity code or set (y-axis) in the forced-choice portion of each game across trials (x-axis). (b, d) The probability that participants eventually solved a game (within 10 guesses) in phase two (y-axis) given the proportion of low-complexity guesses they made in phase one (x-axis). (c, f) The number of guesses participants required to solve the game in phase two (y-axis) given the proportion of low-complexity guesses they made in phase one (x-axis), non-solvers excluded.

freely without any feedback. The options in phase one were designed such that one guess was information-maximizing with no complexity penalty, while the other penalized complexity (as in the two conditions of Experiment 1). Additionally, none of the options presented in the forced-choice phase were the correct code or set. We recruited 50 participants for each game from Prolific, with $n = 40$ remaining in Mastermind and $n = 41$ remaining in ButtonSet after exclusions for not solving at least half the games.

Results

Our main question was whether participants would choose the maximally informative option or the lower-complexity option. Across both games, participants initially picked between the two options near chance but with experience developed a strong preference for the lower-complexity option. In Mastermind, a logistic regression predicting the probability of choosing the low-complexity option as a function of trial showed that the intercept was negative ($B_0 = -0.33$,

$z = -3.04$, $p = 0.002$) and there was a significant positive effect of trial number ($B_t = 0.14$, $z = 6.69$, $p < 0.001$), meaning participants picked the low-complexity option about 44% of the time in the first game but 71% by the final game. In ButtonSet, the intercept was also slightly negative ($B_0 = -0.16$, $z = -1.99$, $p = 0.05$), but participants' increasingly chose the lower-complexity option over time ($B_{\text{trial}} = 0.09$, $z = 4.35$, $p < 0.001$), from 46% in the first game to 66% by the final game (Figure 5a,d).

In both games, there was a striking relationship between selecting lower-complexity guesses and solving the game successfully. In Mastermind, using low-complexity guesses more frequently was a strong predictor of solving the game ($B_{\text{low}} = 1.79$, $z = 4.28$, $p < 0.001$), with participants solving the game only 50% of the time when they always picked the Max-EIG option, but solving 86% of the time when they always picked the low-complexity option. Similarly, in ButtonSet, the proportion of low-complexity guesses also significantly predicted success ($B_{\text{low}} = 1.26$, $z = 3.20$, $p = 0.001$); when participants chose only the Max-EIG option, they solved the game 66% of the time, but when they picked the low-complexity option exclusively, the success rate increased to 87% (Figure 5b,e). Participants also took fewer attempts to solve games when they had chosen more low-complexity queries (Figure 5c,f).

Experiment 3: Free game-play

We finally tested how people played both games without any constraints on what queries they could make or any other manipulation. This allowed us to compare the fit of the bounded-representation model to human performance with three alternatives drawn from prior literature [5, 14, 18, 40, 43, 44]: a noise-free model that always selects the information-maximizing query; a model that makes queries at random from the full space of possible codes until the solution is determined; and a "leaky posterior" model which assumes noisy integration of evidence. Note that the random-querying model here strictly under-performs a "one-and-done" sampling model that draws a single sample from the posterior [44], since the random queries are drawn from the entire space of possible codes rather than from the posterior.

Results

We recruited 50 participants for each game, who played 10 games with 5 generated by each model. After excluding those who solved fewer than half the games, we retained 39 participants (Mastermind) and 42 participants (ButtonSet). In Mastermind, participants found the correct code in 86% of games played. Of the games where participants successfully found the code, they took on average 7.5 queries (SD=2.9). This performance is substantially worse than the information-maximizing model, which takes on average 3.6 guesses (SD=0.5) and at most 4 for any given game. A one-sample t-test showed that the difference in the mean number of queries to solve the game (-3.9) was significant ($t(336) = -24.3$, $p < 0.001$). But, notably, it is also worse than the model that picks a query at random until there is only one possible code remaining, which took on average 4.8 queries (SD=1.0) and a maximum of 7 queries in a sample of 1,000 simulated games. This difference (-2.8) was also significant ($t(336) = -17.7$, $p < 0.001$).

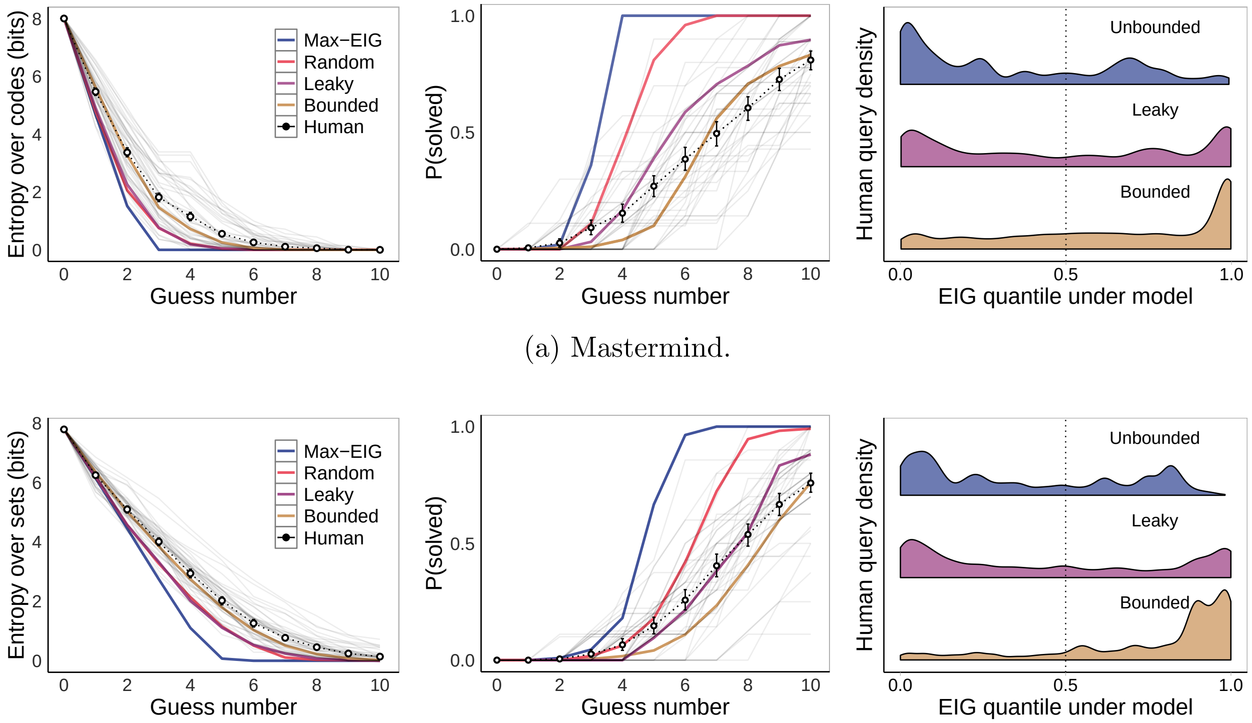
In ButtonSet, participants found the correct solution on 91% of games played and in the games where they found the correct solution, they got the correct answer in 8.5 guesses on average (SD=2.8). As with Mastermind, this is more queries than the unbounded Max-EIG model, which takes 5.1 guesses on average (SD=0.9) and more than the random-querying model, which takes 6.7 guesses on average (SD = 0.9). The difference from the unbounded Max-EIG model was significant ($t(380) = -23.0$, $p < 0.001$) as was the difference from the random model ($t(380) = -12.3$, $p < 0.001$). The uninformativeness of participants' queries and their overall inefficiency in both games can be seen in the lefthand and middle plots in Figure 6 (Mastermind, 6a; ButtonSet, 6b): participants' queries reduce entropy more slowly (left) and they take more queries to find the solution (middle) than both the unbounded Max-EIG model and the random-querying model.

Critically, the fact that the informativeness of participants' queries was below chance in both games cannot be explained by standard sampling-based accounts of approximate inference and decision making [11, 44–49]. Intuitively, the worst these models can perform is if they sample only one possible query or track zero different hypotheses (or "particles"). Either model will select queries completely randomly, and increasing the number of samples will only improve performance. Indeed, we found that a particle filter applied to our tasks selects queries that are more informative than chance (see SI).

We ran a nested model comparison to test the bounded-representation model against an "unbounded" Max-EIG model that assumed only noise in the decision process (i.e., what query to make) and a leaky-posterior model, which additionally assumes noise in inference. In both games, the bounded-representation model provided the best overall fit (average per-participant ΔAIC of 72 for Mastermind and 125 for ButtonSet over the next-best model) and the best fit for a majority of participants. The righthand panel of Figure 6 shows that the modal EIG of participants' queries was close to the (subjectively) information-maximizing query under the bounded-representation model. Participants' queries are in fact objectively uninformative (low average informativity under the unbounded Max-EIG model) and slightly below average under the leaky posterior model. See SI for additional details.

Discussion

In this paper, we proposed and validated a model of how human active learning is shaped by processing demands that formalizes two key hypotheses. First, that people can only approximately represent the implications of data when those implications are complex. Second, that people will adaptively seek out data that might be objectively *uninformative* but will be informative *given their constraints*. Our experiments supported both hypotheses: people could not use highly informative queries effectively (Experiment 1); when given the choice,



(a) Mastermind.

(b) ButtonSet.

Figure 6: Results from Experiment 1, with Mastermind on top (a) and ButtonSet on the bottom (b). Human performance (black) is compared against several models: unbounded information-maximizing (blue), random-querying until the solution is determined (red); leaky integration of evidence (purple); and representationally-bounded inference (tan). Left: entropy over possible codes or sets (y-axis) as a function of the number of guesses taken (x-axis); individual participants are shown in transparent lines. Note that this is the *objective* rather than *subjective* entropy, so it does not account for imperfect inferences. Middle: The number of guesses (y-axis) taken to determine the code or set across all games for participants and the models (x-axis). Right: density estimates for participants' expected information gain relative to other possible guesses under different models: unbounded integration of evidence (blue), a leaky posterior (purple) and a bounded, program-approximated likelihood (tan).

they preferred lower-complexity, lower-information queries (Experiment 2); and their behavior could not be explained by plausible alternative models such as noisy query selection, noisy evidence integration, or particle-based approximate inference (Experiment 3). These findings challenge standard rational accounts of active learning that do not account for processing limitations and support a more nuanced story [cf. 11, 50]: people are cognitively limited in what data they can process effectively; they know this (or can learn their limits quickly); and they make queries with their limited capacity for processing information in mind.

Our results initially seem to contradict well-established principles of human information-seeking based on Expected Information Gain (EIG) and Optimal Experimental Design (OED).

A large body of work has documented instances where adults and children ask questions and seek out data that are likely to be highly informative [3, 8, 51–53]. Furthermore, they can adapt their information-seeking strategies based on environmental statistics, consistent with rational models [7, 18, 42, 52, 54, 55]. Critically, however, these studies employed simple environments in which there are a small number of hypotheses and possible data points, and the relationship between the two is direct. In these cases, representing the full implications of a piece of evidence incurs minimal cost; our model will thus also predict near-optimal information-seeking in such settings. In more complex settings, however, when the hypothesis space is large and evidence is ambiguous, both children and adults have been found to be quite inefficient in absolute terms [10, 12, 18, 56, 57], just as we found in our own experiments. Under our model, the apparent conflict between these two types of findings has a straightforward explanation: when data is easy to reason about, people will approach optimal benchmarks; but when the implications of data are complex, people will struggle to represent those implications, and the efficiency of their information search will suffer accordingly.

Our results add to a growing body of work demonstrating the ways in which human information-seeking deviates from a globally optimal strategy. Other work has demonstrated that human inquiry tends to be myopic, in that people tend to consider only one or a handful of hypotheses at a time [10–12, 58]; they tend to pick questions in a "greedy" fashion rather than choosing questions that are informative with future questions in mind [59, 60]; and their questions are often simple [13, 15, 43]. Our findings extend these results by showing that people exhibit biases in query selection not just due to the number of hypotheses they consider or how far ahead they search, but also due to their inability (or unwillingness) to process the full implications of the available information.

There are several limitations of both our theoretical approach and experimental results. First, our modeling approach does not explain how people actually come up with simplified approximations of evidence. Relatedly, our model does not address at an algorithmic level how people come up with queries that strike a balance between simplicity and informativeness — but our model does suggest one plausible answer. By representing the meaning of evidence in a simple way, people may be able to reason backwards from a desired piece of information (e.g., the number of 2's in the code) to a query that produces this information (e.g., 2222), analogous to the means-ends strategy for problem solving [61].

While our findings highlight limitations in forming complex representations on-the-fly, it's important to consider that these constraints coexist with a remarkable capacity for learning richly structured symbolic systems [62–68]. Returning to the analogy of computer memory architecture, our brains may have severely limited 'RAM' for active processing, but seem to compensate with a vast 'hard drive' of latent representations that can be incrementally built and accessed. Library learning models in program synthesis, which develop increasingly sophisticated abstractions over time by iterative composition and caching, may be a useful framework for understanding how people learn rich structures over longer timescales [69–71].

A promising direction is in integrating realistic cognitive constraints into models of abstraction learning, which account both for people's impoverished ability to actively process information and how these limits may be effectively overcome with better world models.

Methods

Experiment 1

We recruited 50 adult participants from the online platform Prolific to play 10 rounds of Mastermind and ButtonSet (500 games total). They were paid $4 upon completion. This and all other experiments were approved by the University's Institutional Review Board and comply with all relevant ethical regulations. Informed consent was obtained from all participants before beginning the study. All experiments were created using the PsiTurk framework [72].

Games were initialized with 0-3 queries in Mastermind and 0-4 queries in ButtonSet. The initial queries either came from one of 30 games played by the (unbounded) information-maximizing model or the bounded-representation model with $\alpha = 4$. Each participant played five games in each of the two initialization conditions, with the number of initial queries shown selected at random. Participants were instructed to "finish the game from where the computer left off." $N = 50$ participants were recruited from Prolific for both games, 6 of whom were excluded in Mastermind and 5 of whom were excluded from ButtonSet for not solving at least half of the games.

Experiment 2

Experiment 2 was similar to Experiment 1 but had two phases. The first phase was a forced choice, where participants were repeatedly presented two options to select between as a guess, which continued until the code/set was perfectly determined. In phase two, participants tried to guess the code or set freely without any feedback. They had up to 10 guesses to answer correctly, otherwise they lost. The options in the forced-choice phase were designed such that one guess was information-maximizing, while the other was the best guess according to the bounded-representation model with $\alpha = 4$. None of the options presented in the forced-choice phase were the correct code or set. We constructed the choices by dynamically constructing game-trees for 30 games to account for each option a person could choose and what to present next. We recruited 50 participants for each game from Prolific, with $n = 40$ remaining in Mastermind and $n = 41$ remaining in ButtonSet after exclusions for not solving at least half the games.

Experiment 3

On each round of Mastermind, a random code was generated from the 256 possibilities (i.e., 1111, 1112, ... 4444). Players had up to 15 guesses to find the solution, otherwise they lost (see previous section for the rules of Mastermind). Participants who did not find the solution in at least half the games were excluded from analyses ($N = 11$), leaving 39 participants. Our primary analyses also only included trials in which participants found the solution, in order to help exclude motivation as a source of sub-optimal performance.

The methods were nearly identical in ButtonSet, except that instead of an array of boxes to input digits, participants were shown an array of 12 buttons and made queries by clicking a subset (up to 6) of those buttons. As with Mastermind, participants had up to 15 guesses to get the answer correct. We again recruited 50 participants, $N = 8$ of whom were removed for not finding the correct solution in at least half of the games.

Modeling Gameplay

In order for the model not to simply seek information but actually try to win games, we introduced a parameter λ which represents the added utility of guessing the solution. There is also no uncertainty about responses given a query and hypothesis, since these are deterministic. So, the utility function given in (3) becomes,

$$U(q) \propto \lambda \cdot P(q) + \sum_{h \in H} P(h) \cdot D_{KL}\left[P^*(H \mid h, q) \parallel P(H)\right]. \tag{4}$$

where $P(q)$ is the probability that the query is the correct answer. Note that λ modulates the explore-exploit tradeoff, since when λ is low, it will make information-maximizing queries and when λ is high, it will always try to pick the most likely code.

We used a simple "conservative" distance metric [73], which says that an approximation is good to the extent that its extension is similar to the likelihood and it does not rule out possibilities consistent with the evidence. Since the games involve discrete hypotheses with binary likelihoods, we define this as,

$$D\left(P(E \mid \cdot), \pi(\cdot)\right) = \begin{cases} \sum_{h \in H} \left|\pi(h) - P(E \mid h)\right|, & \text{if } \pi(h) \geq P(E \mid h) \text{ for all } h \in H. \\ \infty, & \text{otherwise.} \end{cases} \tag{5}$$

The grammars we used for generating approximations Π, which are shown in Supplementary Tables 1 & 2, are sufficiently expressive that there is always an expression which expresses the full implications of feedback from the game, i.e., where $D\left(P(E \mid \cdot), \pi(\cdot)\right) = 0$. However, these programs can be extremely complicated, particularly when queries involve multiple digits in Mastermind and larger sets of buttons in ButtonSet.

The main challenge to testing this model is that there are almost always a variety of different simplified interpretations of evidence that can be made. For any given piece of information in these games, for example, there are a number of different logical expressions of different complexity which are correspondingly more or less accurate approximations of the information. This is true even at a fixed representational cost α, since there might be equivalently complex and accurate simplifications. It is therefore necessary to marginalize over possible values of both α and the approximations that could be made at each point in a game.

We dealt with this by running the model on each possible guess and feedback one could receive in each game under a range of logarithmically-spaced α parameters (from 1e-4 to 100) and stored the top 10 approximations found from each run of the model (or more if there were ties). For each query/feedback pair, we ran an MCMC-like algorithm to search for approximations, using tree-regeneration proposals from the prior to sample possible approximations with a Metropolis-Hastings acceptance rule, and storing each sample in a list. We ran 4 chains of 100,000 iterations each to form a list of possible approximations and took the top approximations at a given α from this list. We additionally found the *exact* logical expression that matched the semantics of the feedback. All of these were aggregated and used to calculate $P^*(h \mid E, \alpha, \beta)$ for evidence received in a game E, cost parameter α, allowing us to fit the model to human data.

References

1. Popper, K. R. *The Logic of Scientific Discovery* (Routledge, London, 1959).
2. Peirce, C. S. The Fixation of Belief. *Popular Science Monthly* **12,** 1–15 (1877).
3. Oaksford, M. & Chater, N. A rational analysis of the selection task as optimal data selection. *Psychological Review* **101,** 608 (1994).
4. Popper, K. R. Science as falsification. *Conjectures and refutations* **1,** 33–39 (1963).
5. Coenen, A., Nelson, J. D. & Gureckis, T. M. Asking the right questions about the psychology of human inquiry: Nine open challenges. *Psychonomic Bulletin & Review* **26,** 1548–1587 (2019).
6. Nelson, J. D. Finding useful questions: on Bayesian diagnosticity, probability, impact, and information gain. *Psychological review* **112,** 979 (2005).
7. Cook, C., Goodman, N. D. & Schulz, L. E. Where science starts: Spontaneous experiments in preschoolers' exploratory play. *Cognition* **120,** 341–349 (2011).
8. Oaksford, M. & Chater, N. Optimal data selection: Revision, review, and reevaluation. *Psychonomic Bulletin & Review* **10,** 289–318 (2003).
9. Steyvers, M., Tenenbaum, J. B., Wagenmakers, E.-J. & Blum, B. Inferring causal networks from observations and interventions. *Cognitive science* **27,** 453–489 (2003).
10. Markant, D. B., Settles, B. & Gureckis, T. M. Self-directed learning favors local, rather than global, uncertainty. *Cognitive science* **40,** 100–120 (2016).
11. Bramley, N. R., Dayan, P., Griffiths, T. L. & Lagnado, D. A. Formalizing Neurath's ship: Approximate algorithms for online causal learning. *Psychological review* **124,** 301 (2017).
12. Klayman, J. & Ha, Y.-w. Hypothesis testing in rule discovery: Strategy, structure, and content. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **15,** 596 (1989).
13. Grand, G., Pepe, V., Andreas, J. & Tenenbaum, J. B. Loose LIPS Sink Ships: Asking Questions in Battleship with Language-Informed Program Sampling. *arXiv preprint arXiv:2402.19471* (2024).

14. Rothe, A., Lake, B. M. & Gureckis, T. Question asking as program generation. *Advances in neural information processing systems* **30** (2017).

15. Rothe, A., Lake, B. M. & Gureckis, T. M. Do people ask good questions? *Computational Brain & Behavior* **1,** 69–89 (2018).

16. Taylor, N., Hofer, M. & Nelson, J. D. *The paradox of help seeking in the entropy mastermind game* in *Frontiers in education* **5** (2020), 533998.

17. Knuth, D. E. The computer as master mind. *Journal of Recreational Mathematics* **9,** 1–6 (1976).

18. Schulz, E., Bertram, L., Hofer, M. & Nelson, J. D. Exploring the space of human exploration using Entropy Mastermind. *bioRxiv,* 540666 (2019).

19. Miller, G. A. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review* **63,** 81 (1956).

20. Vul, E., Alvarez, G., Tenenbaum, J. & Black, M. Explaining human multiple object tracking as resource-constrained approximate inference in a dynamic probabilistic model. *Advances in neural information processing systems* **22** (2009).

21. Alvarez, G. A. & Franconeri, S. L. How many objects can you track?: Evidence for a resource-limited attentive tracking mechanism. *Journal of vision* **7,** 14–14 (2007).

22. Kaufman, E. L., Lord, M. W., Reese, T. W. & Volkmann, J. The discrimination of visual number. *The American journal of psychology* **62,** 498–525 (1949).

23. Jevons, W. S. The power of numerical discrimination. *Nature* **3,** 281–282 (1871).

24. Luck, S. J. & Vogel, E. K. The capacity of visual working memory for features and conjunctions. *Nature* **390,** 279–281 (1997).

25. Van den Berg, R., Shin, H., Chou, W.-C., George, R. & Ma, W. J. Variability in encoding precision accounts for visual short-term memory limitations. *Proceedings of the National Academy of Sciences* **109,** 8780–8785 (2012).

26. Anderson, J. R. Retrieval of propositional information from long-term memory. *Cognitive psychology* **6,** 451–474 (1974).

27. Anderson, J. R. & Reder, L. M. The fan effect: New results and new theories. *Journal of Experimental Psychology: General* **128,** 186 (1999).

28. Garner, W. R. An informational analysis of absolute judgments of loudness. *Journal of experimental psychology* **46,** 373 (1953).

29. Sims, C. R. Rate–distortion theory and human perception. *Cognition* **152,** 181–198 (2016).

30. Sims, C. R., Jacobs, R. A. & Knill, D. C. An ideal observer analysis of visual working memory. *Psychological review* **119,** 807 (2012).

31. Cheyette, S. J. & Piantadosi, S. T. A unified account of numerosity perception. *Nature Human Behaviour* **4,** 1265–1272 (2020).

32. Russell, R. M. The CRAY-1 computer system. *Communications of the ACM* **21,** 63–72 (1978).

33. Gershman, S. J., Horvitz, E. J. & Tenenbaum, J. B. Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science* **349,** 273–278 (2015).

34. Gigerenzer, G. & Gaissmaier, W. Heuristic decision making. *Annual review of psychology* **62,** 451–482 (2011).

35. Gigerenzer, G. in *Rationality: Psychological and philosophical perspectives* 284–313 (Routledge, 1993).

36. Lieder, F. & Griffiths, T. L. Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and brain sciences* **43,** e1 (2020).

37. Horwich, P. How to choose between empirically indistinguishable theories. *The Journal of Philosophy* **79,** 61–77 (1982).

38. Nussenbaum, K. *et al.* Causal information-seeking strategies change across childhood and adolescence. *Cognitive Science* **44,** e12888 (2020).

39. Gureckis, T. M. & Markant, D. B. Self-directed learning: A cognitive and computational perspective. *Perspectives on Psychological Science* **7,** 464–481 (2012).

40. Callaway, F., Rangel, A. & Griffiths, T. L. Fixation patterns in simple choice reflect optimal information sampling. *PLoS computational biology* **17,** e1008863 (2021).

41. Allen, K. *et al.* Using games to understand the mind. *Nature Human Behaviour,* 1–9 (2024).

42. Coenen, A., Ruggeri, A., Bramley, N. R. & Gureckis, T. M. Testing one or multiple: How beliefs about sparsity affect causal experimentation. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **45,** 1923 (2019).

43. Rothe, A., Lake, B. M. & Gureckis, T. M. *Asking and evaluating natural language questions.* in *CogSci* (2016).

44. Vul, E., Goodman, N., Griffiths, T. L. & Tenenbaum, J. B. One and done? Optimal decisions from very few samples. *Cognitive science* **38,** 599–637 (2014).

45. Griffiths, T. L. & Tenenbaum, J. B. Optimal Predictions in Everyday Cognition. **17,** 767–773 (2006).

46. Sanborn, A. N. & Chater, N. The Sampling Brain. *Trends in Cognitive Sciences* **21,** 492–493. ISSN: 1364-6613, 1879-307X. (2024) (July 2017).

47. Sanborn, A. N., Griffiths, T. L. & Navarro, D. J. Rational Approximations to Rational Models: Alternative Algorithms for Category Learning. *Psychological review* **117,** 1144–1144 (2010).

48. Courville, A. C. & Daw, N. *The Rat as Particle Filter* in *Advances in Neural Information Processing Systems* **20** (Curran Associates, Inc., 2007). (2024).

49. Stewart, N., Chater, N. & Brown, G. D. Decision by sampling. *Cognitive psychology* **53,** 1–26 (2006).

50. Gong, T., Gerstenberg, T., Mayrhofer, R. & Bramley, N. R. Active causal structure learning in continuous time. *Cognitive Psychology* **140,** 101542 (2023).

51. Schulz, L. E., Gopnik, A. & Glymour, C. Preschool children learn about causal structure from conditional interventions. *Developmental science* **10,** 322–332 (2007).

52. Coenen, A., Rehder, B. & Gureckis, T. M. Strategies to intervene on causal systems are adaptively selected. *Cognitive psychology* **79,** 102–133 (2015).

53. Liquin, E. G., Callaway, F. & Lombrozo, T. *Developmental change in what elicits curiosity* in *Proceedings of the annual meeting of the cognitive science society* **43** (2021).

54. Ruggeri, A. & Lombrozo, T. Children adapt their questions to achieve efficient search. *Cognition* **143,** 203–216 (2015).

55. Nelson, J. D., Divjak, B., Gudmundsdottir, G., Martignon, L. F. & Meder, B. Children's sequential information search is sensitive to environmental probabilities. *Cognition* **130,** 74–80 (2014).

56. Chen, Z. & Klahr, D. All other things being equal: Acquisition and transfer of the control of variables strategy. *Child development* **70,** 1098–1120 (1999).

57. Klahr, D., Fay, A. L. & Dunbar, K. Heuristics for scientific experimentation: A developmental study. *Cognitive Psychology* **25,** 111–146 (1993).

58. Gregg, L. W. & Simon, H. A. Process models and stochastic theories of simple concept formation. *Journal of Mathematical Psychology* **4,** 246–276 (1967).

59. Meder, B., Nelson, J. D., Jones, M. & Ruggeri, A. Stepwise versus globally optimal search in children and adults. *Cognition* **191,** 103965 (2019).

60. Bramley, N. R., Lagnado, D. A. & Speekenbrink, M. Conservative forgetful scholars: How people learn causal structure through sequences of interventions. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **41,** 708 (2015).

61. Newell, A., Shaw, J. C. & Simon, H. A. *Report on a general problem solving program* in *IFIP congress* **256** (1959), 64.

62. Lake, B. M., Ullman, T. D., Tenenbaum, J. B. & Gershman, S. J. Building machines that learn and think like people. *Behavioral and brain sciences* **40,** e253 (2017).

63. Tenenbaum, J. B., Kemp, C., Griffiths, T. L. & Goodman, N. D. How to grow a mind: Statistics, structure, and abstraction. *Science* **331,** 1279–1285 (2011).

64. Tenenbaum, J. B., Griffiths, T. L. & Kemp, C. Theory-based Bayesian models of inductive learning and reasoning. *Trends in cognitive sciences* **10,** 309–318 (2006).

65. Kemp, C. & Tenenbaum, J. B. The discovery of structural form. *Proceedings of the National Academy of Sciences* **105,** 10687–10692 (2008).

66. Mills, T., Tenenbaum, J. & Cheyette, S. Human spatiotemporal pattern learning as probabilistic program synthesis. *Advances in Neural Information Processing Systems* **36** (2024).

67. Piantadosi, S. T., Tenenbaum, J. B. & Goodman, N. D. The logical primitives of thought: Empirical foundations for compositional cognitive models. *Psychological review* **123,** 392 (2016).

68. Goodman, N. D., Tenenbaum, J. B., Feldman, J. & Griffiths, T. L. A rational analysis of rule-based concept learning. *Cognitive science* **32,** 108–154 (2008).

69. Zhao, B., Lucas, C. G. & Bramley, N. R. A model of conceptual bootstrapping in human cognition. *Nature Human Behaviour* **8,** 125–136 (2024).

70. Rule, J. S. *et al.* Symbolic metaprogram search improves learning efficiency and explains rule learning in humans. *Nature Communications* **15,** 6847 (2024).

71. Ellis, K. *et al.* DreamCoder: growing generalizable, interpretable knowledge with wake–sleep Bayesian program learning. *Philosophical Transactions of the Royal Society A* **381,** 20220050 (2023).

72. Gureckis, T. M. *et al.* psiTurk: An open-source framework for conducting replicable behavioral experiments online. *Behavior research methods* **48,** 829–842 (2016).

73. Edwards, W. Conservatism in human information processing. *Formal representation of human judgment* (1968).