Prediction, Explanation, and Control Under Free Exploration

Roman Tikhonov (rtikhono@andrew.cmu.edu)
Department of Social & Decision Sciences, Carnegie Mellon University, 5000 Forbes Avenue Pittsburgh, PA 15213 USA

Simon DeDeo (sdedeo@andrew.cmu.edu)
Department of Social & Decision Sciences, Carnegie Mellon University, 5000 Forbes Avenue Pittsburgh, PA 15213 USA
Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA

Abstract

Prediction, explanation, and control are basic cognitive abilities. Here we show how they can arise, simultaneously, from underlying mental models built during unstructured, exploration-based learning. Our experimental paradigm, involving interaction with a symbolic "chatbot", allows us to vary the relative difficulty of the tasks, and to measure how participants leverage the Bayesian evidence of their mental models for decision-making. Our experimental manipulation focuses on hidden information and task complexity. With full information, there are significant differences between the three tasks: for example, people are more sensitive to Bayesian evidence in prediction than in control or explanation. When information is hidden, however, performance equalizes. Taken together, our results suggest that, while specific heuristics may lead to different levels of performance in cases with full information, more fundamental forms of reasoning, based on an underlying mental model, and less sensitive to the specific task, come into play when pieces are missing.

Keywords: exploration-based learning; explanation; prediction; control; counterfactual reasoning; finite-state machines; dynamic decision-making

Introduction

Prediction, explanation, and control are computationally distinct cognitive abilities that are usually studied in isolation (e.g., Griffiths & Tenenbaum, 2009; Bubic, Von Cramon, & Schubotz, 2010; Horne, Muradoglu, & Cimpian, 2019; Osman, 2010; Uppal, Ferdinand, & Marzen, 2020; Gerstenberg, Goodman, Lagnado, & Tenenbaum, 2021), but they are a core trio of tasks that, in the real world, are often called upon in rapid succession. Driving on a highway late at night, for example, we might first try to predict what an oncoming car will do; explain why its behavior is out of the ordinary; then control the outcome by flashing our lights, or honking our horn, to avoid an accident.

Theoretical accounts of human learning typically put forward one of these three abilities, giving the other two subordinate roles. *Prediction-first* theories (Friston, 2010; Friston et al., 2015; Hohwy, 2013; Clark, 2013) assume that the ultimate goal of the human mind is to minimize the error between predicted and actual inputs, while our abilities to explain and control emerge as the result of this prediction-driven activity. In contrast, proponents of the *explanation-first* approach (Lombrozo, 2006; Byrne, 2016; Wojtowicz & DeDeo, 2020) suggest that people are driven by the desire to build an accurate model of the causal structure of their environment, which serves as the basis for their further decisions and predictions. *Control-first* frameworks, on the other hand, describe how the ability to control the environment can arise from mechanisms that do not necessarily involve prediction or explanation—e.g., instance-based learning (Gonzalez, Lerch, & Lebiere, 2003), reinforcement learning (Silver, Singh, Precup, & Sutton, 2021), or heuristic decision-making (Gigerenzer, 2001). While each of these approaches offers strong assumptions about the relationship between prediction, explanation, and control, there is a lack of empirical studies that simultaneously examine all three abilities, making it difficult to draw clear conclusions in favor of one approach over another.

While each of these theories implies a hierarchical relationship between prediction, explanation, and control—with one of these implicitly driving the others—isolated studies suggest that these abilities have complex relationships to each other. For example, Fernbach, Darlow, and Sloman (2010, 2011) showed that people are better at making diagnostic (i.e., explanatory, backward-reasoning) judgments compared to predictive, forward-reasoning judgments about the probabilities of future events. They referred to it as an *alternative neglect bias*, which is a tendency to ignore alternative causes of a given event. Additionally, studies of human performance in dynamic system control tasks have found that people can be equally good at control and prediction under a salient rule, but when the pattern becomes less obvious, people can still control the system but cannot predict it (Berry & Broadbent, 1984, 1988).

Fundamentally, prediction, explanation, and control are goal-oriented tasks that involve finding the "right" answer. However, individuals often learn to perform these tasks through a period of self-guided free exploration in the absence of strong goals. This period is typically guided by epistemic drives such as curiosity (e.g., Dubey, Mehta, & Lombrozo, 2021) or belief-based utility (Golman & Loewenstein, 2018). While an adult at a cocktail party may be able to explain why, for example, a companion is upset, or be able to find the right words to support them, these talents are honed through years of experience in simply talking with others, without explicit training.

This experience—of an early period of "playful" interaction leading to good, and even expert, performance on goal-oriented metrics—may well be the dominant form of learning in childhood (Gopnik, 2020). Even in adult life (Gottlieb & Oudeyer, 2018) we interact with a world where explicit feed-

In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society.*

Task	Basic Form	Question Format	Normative Answer
Prediction	*What is the most likely next state?*	Visible: $1_a 2_b \underline{?}$	$\text{argmax}_i P(i\|2_b)$
		Hidden: $1_a \text{X}_b \underline{?}$	$\text{argmax}_i \left(\sum_{k=1}^{N} P(i\|k_b)P(k\|1_a)\right)$
Control	*What choice of inputs is most likely to put the system into the goal state?*	Visible: $1_{\underline{a}} 2_{\underline{?}} 3$	$\text{argmax}_j P(3\|2_j)$
		Hidden: $1_{\underline{?}} \text{X}_{\underline{?}} 3$	$\text{argmax}_{\{i,j\}} \left(\sum_{k=1}^{N} P(3\|k_j)P(k\|1_i)\right)$
Explanation	*Which input caused the system to be in that particular final state?*	Visible: $1_{\underline{a}} 2_{\underline{b}} 3$	$P(3\|2_b)P(2\|1_b) < P(3\|2_a)P(2\|1_a)$
		Hidden: $1_{\underline{a}} \text{X}_{\underline{b}} 3$	$\sum_{k=1}^{N} P(3\|k_b)P(k\|1_b) < \sum_{k=1}^{N} P(3\|k_a)P(k\|1_a)$

Note. Here, N is the number of intermediate states. The probabilities P are given by the subject's mental model, which is learned from free exploration in phase one of the experiment. In the explanation task, we show the case where the correct answer is "because the first input was a".

Table 1: Normative answers to the test tasks. States are digits from 1 to 4, responses are subscript lowercase letters (a, b), hidden information is X, and queries are "?" (see also Fig. 1)

back on performance is rare and ambiguous. This paper seeks to understand the relationship between these two stages. We are interested in both how someone constructs a mental model of a system through undirected interaction without specific goals or incentives, and then how they use that mental model when called upon to do explicit prediction, explanation, and control.

Our study aims at answering the following questions: (1) Could one learn to predict, control, and explain a dynamic system via free exploration? (2) How would the performance change in the presence of hidden information, when multiple unobserved possibilities had to be considered? (3) How do these answers vary with the complexity of the underlying system? To answer these questions, we conduct an experiment where participants first interact with a dynamic system that follows a set of rules (see Fig. 1 and Fig. 2), and then answer a set of questions designed to evaluate their ability to predict, explain, or control the dynamical system (Fig. 3). Additionally, we vary the amount of information provided with each test question by hiding or uncovering some pieces of information that are necessary to make a correct judgment.

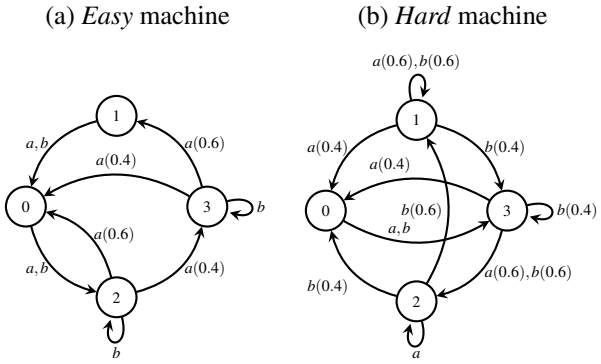

Figure 1: Finite-state machines (FSMs) that define responses of the *chatbot*. Each of the four states represents an emoji icon sent by the chatbot, while inputs (a, b) correspond to emojis sent by participants. Parentheses indicate probabilities for inputs with multiple next states.

A Model of Mental Models

Drawing on recent work (Tikhonov, Marzen, & DeDeo, 2022), we understand prediction, explanation, and control tasks as relying on an underlying mental model of the system—a largely tacit, implicit, and probabilistic representation of how the system works. In our interpretation of the experiments below, the idea is that a subject constructs a mental model during the free exploration phase, which is then used to answer the test questions in the second phase.

This is challenging. An individual who possesses a mental model will, in general, have only partial access to its structure and can only articulate some fraction of what it contains. Called upon to predict, explain, or control a system in response to test questions, the individual faces the challenge of making some aspects of the model explicit, and capable of guided deliberative action.

We model this translation of implicit knowledge to explicit decision-making in two stages. In the first stage, we consider the participant's mental model of the underlying machine. This takes a Bayesian form, specifying the (probabilistic) response of the machine to different inputs: explicitly, the mental model gives the reasoner access to the probability distribution $P(i|j_k)$, where j is the current state, k is the current input symbol, and i is the next state.

This model is then used by the agent to judge the relative likelihoods for the different tasks. The form of the tasks and the normative answers—*i.e.*, the way in which the correct answers are calculated—are shown in Table 1. In the prediction case, for example, the participant is asked to predict the response of the machine to a sequence of inputs, making a binary choice between final state A (say), and final state B. The mental model provides a degree of belief in these two outcomes, $P(A)$, and $P(B)$, which can be summarized as the relative log-likelihood, R, of the more likely choice; if $P(A)$ is larger than $P(B)$, this is

$$R = \log \frac{P(A)+\varepsilon}{P(B)+\varepsilon}, \tag{1}$$

where ε is a small regularizing parameter that takes into ac-

count that the participant may attribute some small probability to an outcome that their mental model says is, formally, impossible.

The value of R is taken to be more-or-less implicit content, which the participant needs to act on. We assume that this happens in a noisy fashion; if A is the correct answer at evidence level R, then the participant chooses A with probability p_C given by

$$p_C = \frac{\exp(\beta R)}{\exp(\beta R) + \exp(-\beta R)}, \quad (2)$$

where β parameterizes the noise in the translation from implicit to explicit. When β is large, the participant makes efficient use of the knowledge R; when it is small, the choice is much less reliable; when it is equal to zero, the choice is random.

To use this model to understand human decision-making, we first construct an approximation to the mental model we believe the participant possesses on the basis of their free exploration, based upon a simple frequency-based rule: $P(i|j_k)$ is equal to the number of times the participant observes state i follow state j under input k, divided by the number of times they saw state j under input k. For any question i, we then compute the relative probabilities, R_i, of the two options for the answer to a task question (see Table 1): in the prediction task, this might be the relative probability of the system ending up in State 1 versus State 2; in the control task, the relative probability of the system ending up in the desired state, given that the agent chooses to do either action a or action b; in the causal (counterfactual) explanation task, the relative probability that the system would have behaved differently if action a was not done, versus action b. The is given by Eq. 1.

Finally, we see how well the choice indicated by the mental model matches the actual behavior of the participant. Formally, this is simple: the β parameter of such a model is simply the coefficient in a logistic regression on the correct answer to question i, with independent variable R_i, without intercept.

The reason for this rather elaborate process is that different tasks will have different difficulties: a person's mental model may give clear guidance for a prediction task (*i.e.*, suggest a decisive choice, with large R_i), but a much weaker one for a control task. What we care about is the reliability of the *use* of the model at a fixed level of evidence (given by β), not the actual performance, which is a mixture of both β and R. If we simply score participants on performance, we will confuse tasks that are difficult because the answers are less clear, with tasks that are difficult because participants struggle to use their mental model well.

A second benefit of our approach is that it allows us to compare different ways a participant might use a mental model. While this paper uses only the normative account of Table 1 to define what is meant by prediction, explanation, and control, it is possible, as discussed in Tikhonov et al. (2022) to consider an alternative, non-normative, forms of the three tasks—"alternative neglect"—to see if there is evidence for the use of this distinct heuristic.

Methods

With the analysis procedure, above, in hand, we applied it in an online experiment.

Participants

Ninety-seven English-speaking U.S. participants (49 men and 48 women; 18-47 y.o., $M_{age} = 27.3, SD_{age} = 6.7$) with normal or corrected to normal vision, were recruited online via Prolific for a $2 compensation with a performance-based bonus up to $2. The study took approximately nine minutes and required a desktop or laptop computer.

Materials and Procedure[1]

Dynamic Interaction Task We developed an experimental paradigm that was modeled after Berry and Broadbent's (1984) *Personal Interaction Task*, originally designed to investigate implicit and explicit knowledge in dynamic system control. In our study, participants interact with a "chatbot" using a fixed set of emoji icons. The procedure includes a learning phase (45 interactions), a test phase (20 trials), and a short questionnaire.

The chatbot's behavior is defined by a finite-state machine (FSM) with four states and transitions between them guided by two inputs. Participants are randomly assigned to a condition associated with one of two machines (see Fig. 1) that differed in the number of probabilistic and deterministic transitions. The *easy* machine has two probabilistic and six deterministic transitions. The *hard* machine has five probabilistic transitions and three deterministic, so it requires much more effort to be learned.

Learning Phase At the beginning of the learning phase, participants were told that the chatbot's responses follow a certain pattern and were asked to freely interact with the chatbot "to get a sense of how it responds to different messages so that they would be able to explain, predict, and control its behavior." They were also asked not to use any outside resources or assistance. The chatbot begins in a random state (see Fig. 1), emitting the associated emoji. Participants respond by choosing one of the two emoji icons (corresponding to a, or b) and instantly get the next reaction of the chatbot, which depends on their input and the previous message from the chatbot (Fig. 2).

Test Phase Participants were randomly assigned to one of three test conditions that assessed their ability to predict, explain, or control the chatbot's behavior. All test tasks were presented as episodes of conversation with the chatbot (State_1–Input_1–State_2–Input_2–State_3), exactly paralleling the format of Table 1, with a two-alternative forced choice question corresponding to the test condition. As a within-subjects

[1]Materials, data, and analysis scripts are available at https://osf.io/59m8r/

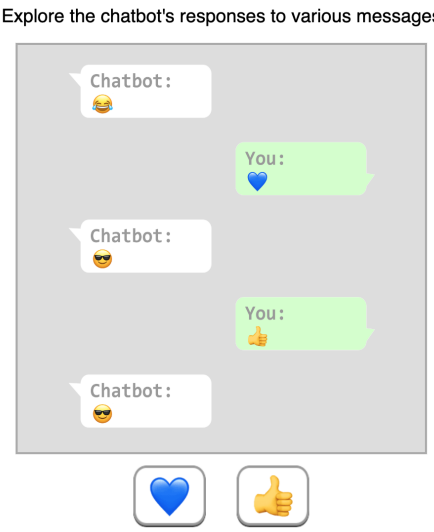

Figure 2: Interaction with the chatbot at the learning ("free exploration") phase.

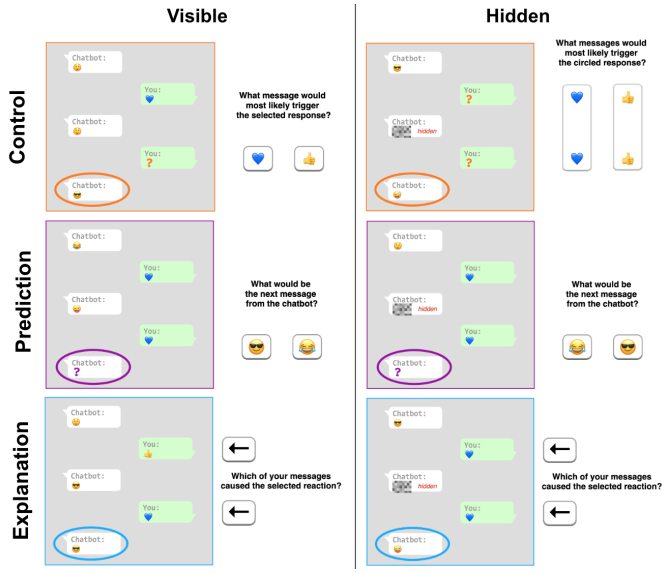

Figure 3: Examples of questions presented during the test phase. Participants selected their answers by clicking on the corresponding buttons.

variable, a form of test question (*visible* or *hidden*) was manipulated by presenting or hiding an intermediate message from the chatbot (see Fig. 3). A total of 20 questions—ten hidden and ten visible—were asked in random order. Participants received 10 cents for each correct answer as a bonus payment.

The goal of the **prediction** task was to see if participants could correctly anticipate the chatbot's next response by looking at past interactions. Visible form included a State_1–Input_1–State_2–Input_2–[Question] combination along with a question (*What would be the next message from the chatbot?*) and two states as answer options. The hidden form was identical except for State_2 being concealed.

Visible **control** tasks were presented as State_1–Input_1–State_2–[Question]–State_3 episodes with a question (*What messages would most likely trigger the selected response?*) and two inputs (*a* or *b*) as answer options. Hidden control tasks contained only State_1 and State_3 with State_2 and both inputs being hidden. Answer options included two (out of four possible) combinations of Input_1 and Input_2. Participants had to determine which message or combination of messages would evoke a specific response from the chatbot.

In the **explanation** condition, the task was to decide which of the previous messages caused the chatbot's final reaction. In the instructions, we emphasize that the message that causes the final response need not be the one that occurred immediately before: "It can sometimes be the case, for example, that once a certain action is taken, the next action has little or no ability to change the outcome. In this case, the earlier action may have been the cause." Conversation episodes were presented as State_1–Input_1–State_2–Input_2–State_3 with State_2 being visible or hidden along with a question (*Which of your messages caused the selected reaction?*) and two answer options—buttons pointing at Input_1 or Input_2.

Results

We evaluated participants' learning performance using two accuracy measures. *Actual accuracy* relied on FSMs as the ground truth to determine the proportion of correct answers, while *mental model accuracy* was more relative, based on mental models inferred from probabilities observed during self-guided free exploration.[2] Figure 4 shows that actual and mental model accuracies produced almost identical results, and a paired-samples t-test found no significant differences between them (mean accuracies were 60% and $SDs = .18$; $t(193) = 0.90$, $p = .370$). Participants were able to predict, control, and explain both visible and hidden forms of questions of the *easy* FSM. However, in the *hard* FSM condition, only control remained at the above chance level (see Figure 4). As our primary interest is in understanding how participants build and apply their mental models, we focused on mental model accuracy in our analysis.

We conducted pairwise t-tests with Benjamini-Hochberg adjustment for multiple comparisons to examine the differences between test conditions. In the *easy* FSM condition, participants performed better in prediction ($M = 0.81$, $SD = 0.13$) than in explanation ($M = 0.58$, $SD = 0.14$) when responding to the visible form of questions ($t(30.7) = 4.83$, $p < .001$), but there was no statistically significant difference in the hidden form ($t(30.7) = 0.46$, $p = .891$). No other statistically significant differences were found in the *easy* FSM.

[2] In rare cases where a state-input combination had never occurred in the learning phase, we assigned equal probabilities to all four next states.

In the *hard* FSM condition, there was a performance advantage in control ($M = 0.66$, $SD = 0.10$) compared to prediction ($M = 0.46$, $SD = 0.18$, $t(30) = 3.83$, $p = .004$) and explanation ($M = 0.55$, $SD = 0.13$, $t(29.9) = 2.80$, $p = .026$) for the visible form of questions. In the hidden form of questions, we found no significant differences. Overall, participant performance varied across prediction, explanation, and control tasks in the visible form of questions, with prediction superior in the easy FSM condition and control performing better in the hard FSM condition. However, when the intermediate state was hidden, performance on prediction, explanation, and control tasks became similar.

To explore how participants applied their mental models to predict, explain, and control dynamic systems, we conducted logistic regression analysis. We used the Bayesian evidence of answer choices (R_i) and its interaction with question form (visible/hidden) as predictors of mental model accuracy. Larger R_i values correspond to questions that should be easier to answer assuming that participants interact with their mental models in a Bayes-like fashion. As indicated in Table , the answer choice evidence (R_i) was a statistically significant predictor of accuracy across tasks in the *easy* FSM condition, suggesting that participants generally used evidence in a Bayesian manner. When the intermediate state was hidden, the interaction between R_i and question form became a negative predictor of accuracy in control and prediction tasks, suggesting that limited information availability might impede the efficient use of mental models. In the *hard* FSM condition, R_i remained a positive statistically significant predictor of accuracy only for control, suggesting that machine complexity influenced the extent of Bayesian information use.

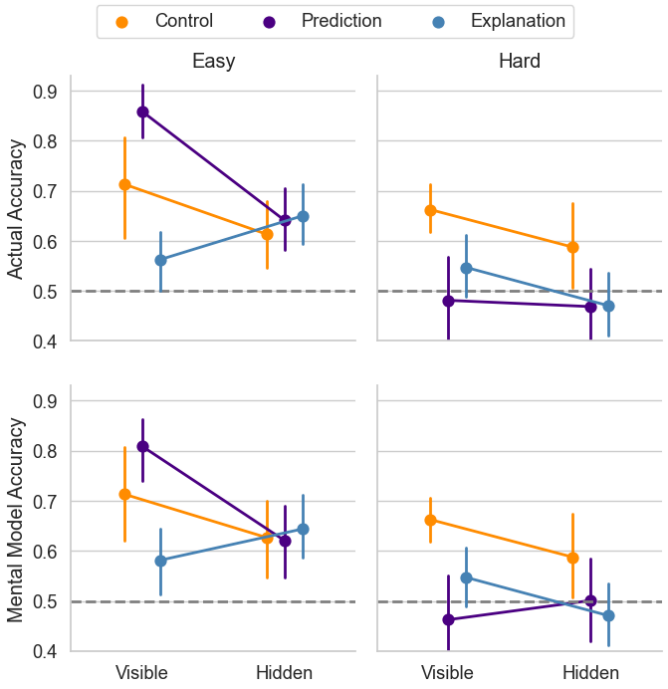

Figure 4: Actual and Mental Model Accuracy. Means and 95% CIs were calculated from aggregated participant data.

Discussion

In the course of day-to-day life, we are sometimes tested, more or less explicitly, on our abilities to predict, explain, and control. How we perform on these occasional tests—and the "feedback" we receive—can directly impact our flourishing, if not our chances of survival. Such tests, however, with their immediate feedback, are relatively rare. The learning that enables us to perform well happens under very different circumstances. What enables us to survive is often the product of many years of experience with no tests at all—there is a gap, in other words, between the things we do to gain the ability, and the way in which those abilities are tested.

This paper has taken that gap seriously. Instead of seeing how people train on a task in the presence of feedback, we first present them with a system to explore in an unstructured fashion. To make sense of their subsequent performance, we then think of them as relying on the mental model they constructed in the first phase, and examine the extent to which they are able to leverage that learning for what comes next.

Our most basic finding is clear: individuals can, indeed, successfully learn to predict, explain, and control simple dynamic systems through self-guided free exploration; they demonstrate not only significantly above-chance performance in all three tasks, but also a clear relationship between performance on a question, and the Bayesian evidence that would be present where they using mental models in the way we expect. These results are robust even when they encounter missing information and have to navigate through multiple possibilities. Interestingly, we find that some tasks are more "Bayesian" than others; prediction, in particular, is more sensitive to Bayesian evidence than control and explanation.

In the simplest cases, prediction performance is much better than explanation, with control performance falling in between. Explanation may bring many pleasures (Gopnik, 1998), but it appears more difficult to achieve than the more prosaic tasks. This is, on the face of it, counterintuitive: the ability to control, say, would seem to require mastery of, informally, "what causes what", and thus some ability to explain. Our results, however, suggest that—at least in the presence of complete information—good-enough performance can be achieved even in the absence of a causal model. A natural explanation for this phenomenon is that control abilities (for example) can be gained by heuristics and memorization, without needing to rely on a more complex model of the world that would, indeed, give explanatory abilities.

In favor of this account is the fact that these large differences exist only in the presence of complete information. The differences in performance disappear when information is hidden. This suggests that, when information is hidden, participants are no longer able to rely on task-specific heuristics, and must fall back, instead, on the use of the underlying

Task	$\beta^{\text{Upper 95\% CI}}_{\text{Lower 95\% CI}}$	z	p
Easy FSM			
Prediction			
R_i	$0.90\ ^{1.14}_{0.68}$	7.69	$< .001$***
$R_i \times$ Hidden	$-0.51^{-0.19}_{-0.83}$	-3.10	.002**
Control			
R_i	$0.48\ ^{0.67}_{0.31}$	5.26	$< .001$***
$R_i \times$ Hidden	$-0.27^{-0.01}_{-0.54}$	-2.04	.041*
Explanation			
R_i	$0.18\ ^{0.35}_{0.01}$	2.10	.036*
$R_i \times$ Hidden	$0.17\ ^{0.42}_{-0.08}$	1.31	.189
Hard FSM			
Prediction			
R_i	$-0.06\ ^{0.10}_{-0.22}$	-0.68	.494
$R_i \times$ Hidden	$0.14\ ^{0.39}_{-0.11}$	1.07	.285
Control			
R_i	$0.40\ ^{0.58}_{0.23}$	4.40	$< .001$***
$R_i \times$ Hidden	$-0.20\ ^{0.08}_{-0.47}$	-1.42	.156
Explanation			
R_i	$0.18\ ^{0.39}_{-0.01}$	1.83	.068
$R_i \times$ Hidden	$-0.19\ ^{0.09}_{-0.49}$	-1.32	.188

Note. The intercept was excluded to facilitate the interpretation of the R_i effects.
* $p \leq .05$, ** $p \leq .01$, *** $p \leq .001$.

Table 2: Predicting accuracy in different test conditions based on answer choice evidence (R_i) and its interactions with question form.

mental model. When this happens, we expect performance to equalize simply because, as can be seen from Table 1, the complexity of the three tasks—everything from the number of things to keep track of, to the number of hidden states to marginalize over—is identical, and the only difference is the order in which the terms are multiplied and summed (a minor difference is that control requires an argmax over two variables, but this is compensated by the fact that the decision process is always over a binary forced choice). According to our theoretical approach, when people rely upon probabilistic mental models, in other words, they ought to produce similar performance across all three tasks, and this is precisely what we find in the case of hidden information.

Returning to the full information case, a third finding of our work is that control is more robust to information complexity than prediction and explanation. Specifically, participants were able to maintain their performance in control tasks even under the hard FSM condition, while their ability to predict and explain the system deteriorated. This is consistent with previous research on implicit learning of dynamic systems (Berry & Broadbent, 1984, 1988). However, we note that this finding may have been influenced by the fact that free exploration learning may have been more similar to the control task, as participants had to actively choose inputs in both cases.

A major limitation in our work is given by the large heterogeneity in performance. Looking at the individual level, we find that, in many cases, some fraction of our participants performed no better than random, while others achieved near-perfect accuracy. (Our limits on response time rule out the possibility that top performers are, for example, building explicit models using pencil and paper.) A minority of participants—around 20%—can achieve excellent performance on the very hardest tasks, well above chance: 80% to 90% accuracy, for example, in causal reasoning about complex machines under partial information.

Acknowledgements

This work was supported in part by the Survival and Flourishing Fund. We thank our reviewers, meta-reviewer, and Victor Møller-Poulsen for their very helpful remarks.

References

Berry, D. C., & Broadbent, D. E. (1984). On the relationship between task performance and associated verbalizable knowledge. *The Quarterly Journal of Experimental Psychology Section A*, *36*(2), 209-231. doi: 10.1080/14640748408402156

Berry, D. C., & Broadbent, D. E. (1988). Interactive tasks and the implicit-explicit distinction. *British Journal of Psychology*, *79*(2), 251–272. doi: 10.1111/j.2044-8295.1988.tb02286.x

Bubic, A., Von Cramon, D. Y., & Schubotz, R. (2010). Prediction, cognition and the brain. *Frontiers in Human Neuroscience*, *4*. doi: 10.3389/fnhum.2010.00025

Byrne, R. M. (2016). Counterfactual Thought. *Annual Review of Psychology*, *67*(1), 135–157. doi: 10.1146/annurev-psych-122414-033249

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*(3), 181–204. (Publisher: Cambridge University Press) doi: 10.1017/S0140525X12000477

Dubey, R., Mehta, H., & Lombrozo, T. (2021). Curiosity Is Contagious: A Social Influence Intervention to Induce Curiosity. *Cognitive Science*, *45*(2), e12937. doi: 10.1111/cogs.12937

Fernbach, P. M., Darlow, A., & Sloman, S. A. (2010). Neglect of Alternative Causes in Predictive but Not Diagnostic Reasoning. *Psychological Science*, *21*(3), 329–336. (Publisher: SAGE Publications Inc) doi: 10.1177/0956797610361430

Fernbach, P. M., Darlow, A., & Sloman, S. A. (2011). Asymmetries in predictive and diagnostic reasoning. *Journal of Experimental Psychology: General*, *140*(2), 168–185. (Publisher: American Psychological Association) doi: 10.1037/a0022100

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, *11*(2), 127–138. doi: 10.1038/nrn2787

Friston, K., Rigoli, F., Ognibene, D., Mathys, C., Fitzgerald, T., & Pezzulo, G. (2015). Active inference and epistemic value. *Cognitive Neuroscience*, *6*(4), 187–214. doi: 10.1080/17588928.2015.1020053

Gerstenberg, T., Goodman, N. D., Lagnado, D. A., & Tenenbaum, J. B. (2021). A counterfactual simulation model of causal judgments for physical events. *Psychological Review*, *128*(5), 936–975. doi: 10.1037/rev0000281

Gigerenzer, G. (2001). The adaptive toolbox. In *Bounded rationality: The adaptive toolbox* (pp. 37–50). Cambridge, MA, US: The MIT Press.

Golman, R., & Loewenstein, G. (2018). Information gaps: A theory of preferences regarding the presence and absence of information. *Decision*, *5*(3), 143.

Gonzalez, C., Lerch, J. F., & Lebiere, C. (2003). Instance-based learning in dynamic decision making. *Cognitive Science*, *27*(4), 591–635. doi: 10.1207/s15516709cog2704$_2$

Gopnik, A. (1998). Explanation as orgasm. *Minds and Machines*, *8*, 101–118.

Gopnik, A. (2020). Childhood as a solution to explore–exploit tensions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *375*(1803), 20190502. doi: 10.1098/rstb.2019.0502

Gottlieb, J., & Oudeyer, P.-Y. (2018). Towards a neuroscience of active sampling and curiosity. *Nature Reviews Neuroscience*, *19*(12), 758–770. (Number: 12 Publisher: Nature Publishing Group) doi: 10.1038/s41583-018-0078-0

Griffiths, T. L., & Tenenbaum, J. B. (2009). Theory-Based Causal Induction. *Psychological Review*, *116*(4), 661–716. (Publisher: American Psychological Association (APA)) doi: 10.1037/a0017201

Hohwy, J. (2013). *The predictive mind*. Oxford University Press.

Horne, Z., Muradoglu, M., & Cimpian, A. (2019). Explanation as a Cognitive Process. *Trends in Cognitive Sciences*, *23*(3), 187–199. doi: 10.1016/j.tics.2018.12.004

Lombrozo, T. (2006). The structure and function of explanations. *Trends in Cognitive Sciences*, *10*(10), 464–470. doi: 10.1016/j.tics.2006.08.004

Osman, M. (2010). Controlling uncertainty: A review of human behavior in complex dynamic environments. *Psychological Bulletin*, *136*(1), 65–86. doi: 10.1037/a0017815

Silver, D., Singh, S., Precup, D., & Sutton, R. S. (2021). Reward is enough. *Artificial Intelligence*, *299*, 103535. doi: 10.1016/j.artint.2021.103535

Tikhonov, R., Marzen, S., & DeDeo, S. (2022). How Predictive Minds Explain and Control Dynamical Systems. In *NeurIPS 2022 Workshop on Information-Theoretic Principles in Cognitive Systems.* Retrieved from `https://openreview.net/forum?id=xk41NgCFxrj`

Uppal, A., Ferdinand, V., & Marzen, S. (2020). Inferring an Observer's Prediction Strategy in Sequence Learning Experiments. *Entropy*, *22*(8), 896. doi: 10.3390/e22080896

Wojtowicz, Z., & DeDeo, S. (2020). From Probability to Consilience: How Explanatory Values Implement Bayesian Reasoning. *Trends in Cognitive Sciences*, *24*(12), 981–993. doi: 10.1016/j.tics.2020.09.013