Prediction, Explanation, and Control Under Free Exploration

Roman Tikhonov (rtikhono@andrew.cmu.edu)[1]

Simon DeDeo (sdedeo@andrew.cmu.edu)[1, 2]

[1]Department of Social & Decision Sciences, Carnegie Mellon University,
5000 Forbes Avenue Pittsburgh, PA 15213 USA

[2]Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA

Abstract

Prediction, explanation, and control are basic cognitive tasks. Here we show how they can arise from unstructured exploration-based learning. We present a new method to analyze how people make use of mental models learned under free exploration. Our method allows us to control for the relative difficulty of the task, and to measure the extent to which participants can leverage the evidence provided by the mental model—however strong or weak—for decision-making. The key result of our subsequent experimental work is that free exploration leads, in simple cases, to far better performance on prediction and control compared to explanation. As systems become more complex, performance declines, but people can sustain relatively better performance on control. In the presence of hidden variables, explanatory abilities rise relative to the other two, even though the task is harder.

Keywords: exploration-based learning; finite-state machines; dynamic decision-making; chatbot interaction task

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 explanation and control serve a secondary purpose. 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 describe how the ability to control the environment (i.e., take actions that help one achieve their goal) can arise from mechanisms that do not necessarily imply 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). All three approaches have rather strong assumptions about the relationship between prediction, explanation, and control. However, the lack of empirical studies that simultaneously examine these abilities makes it difficult to draw a clear conclusion in favor of one approach or another.

Even though prediction- and explanation-first theories imply hierarchical relationships between prediction, explanation, and control, isolated studies suggest that these abilities are not necessarily linked. 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. Another line of research by Berry and Broadbent (1984, 1988) looked at human performance in a dynamic system control task along with their ability to predict it measured with a subsequent questionnaire. They found that people can be equally good at control and prediction under a salient rule. When the pattern becomes less obvious, people could still control the dynamic system, but cannot predict it.

Fundamentally, while prediction, explanation, and control are goal-oriented tasks—in each case, there is a "right" answer one strives to get—one often learns to do them through a period of self-guided free exploration, where one interacts with a system in the absence of strong goals. A video game player, for example, may gain an edge over an opponent by correct prediction of an outcome, or expert control over a weapon, but these skills are learned, in part, in an early exploratory phase where the player simply "plays", experimenting with the system without a strong drive to, say, maximize her score and, potentially, under the influence of epistemic drives such as curiosity (e.g., Dubey, Mehta, & Lombrozo, 2021) or belief-based utility (Golman & Loewenstein, 2018). An adult at a cocktail party may be able to explain why, for example, a companion is upset, but these talents are honed through years of experience in simply talking with others, without explicit training against an explanatory benchmark.

Indeed, 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 feedback 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.

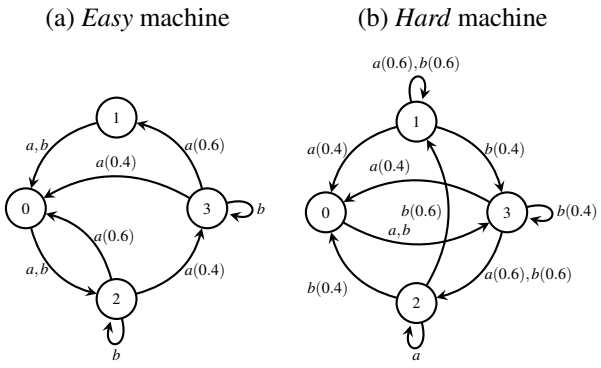

Figure 1: Finite-state machines that define responses of the *chatbot*. Each of the four states represents an emoji icon sent by the chatbot, while inputs (a, b) correspond to emojis sent by participants. Parentheses indicate probabilities for inputs with multiple next states.

Our study aims at answering the following questions: (1) Could one learn to predict, control, and explain a dynamical system via free exploration? (2) How would the performance change under uncertainty, when multiple possibilities had to be considered? (3) How does the structure of a dynamic system affect one's ability to predict, explain, and control it? 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 then answer a set of questions designed to evaluate their ability to predict, explain, or control the dynamical system. Additionally, we vary the level of uncertainty associated with test questions by hiding or uncovering some pieces of information that are necessary to make a correct judgment.

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 the free exploration of subjects leads to the construction of those mental models that are then used to answer the test questions in the second phase.

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. This model is then used by the agent to judge the relative likelihoods for the different tasks.

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 account 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

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

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.

Taken together, this model allows us to translate the degree of evidence provided by the underlying mental model, R, into the test response, whose error is parametrized by an "inverse temperature", β. When R gets larger, there is stronger evidence in favor of one option over the other, making the decision easier; this can be offset by a small value of β, however, indicating that the evidence can not be used to guide action. Crucially, the answer the participant gives need not be "correct, given reality": rather, it must be "correct, given the mental model"; depending on the free exploration stage, a participant may not build a good mental model. (Practically, in our data, there is little difference, but the distinction matters at the theoretical.)

This approach allows us to do a Bayesian inference on actual performance in an experiment with participants. This happens in two stages. First, we construct an approximation

to the mental model we believe the participant possesses on the basis of their free exploration. Then, we compute the relative probabilities they attribute to the system behaving in a certain way; in the prediction task, this is 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*. This is given by Eq. 1.

Finally, we see how well the choice indicated by the mental model matches the actual behavior of the participant. This is done in a Bayesian, maximum-likelihood fashion: we find the value of β in Eq. 2 that best predicts the actual choices. For example, if the person is guessing randomly, β will be around zero; as they get closer and closer to perfect performance, β gets larger and larger.

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), but a much weaker one for a control task. What we care about is the reliability of the *use* of the model at 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 we rely on the normative account of Tikhonov et al. (2022) to define what is meant by prediction, explanation, and control, we also include their second, non-normative, form of causal explanation—"explanation with alternative neglect"—to see if there is evidence for the use of this distinct heuristic.

The β parameter estimates how well the participant can use the mental model, but it is somewhat obscure. In the results presented below, we present our estimates of participant βs in terms of "predicted performance on a standard task with R equal to 3" ("predicted performance" in tables and figures). This presents the results in terms of expected performance at a fixed level of evidence from the mental model; it is a number between 50% (β equal to zero, participant has no access to the mental model, and must guess) and 100% (β very large, participant uses the mental model perfectly).

Methods

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

Participants

Ninety-seven English-speaking U.S. participants (44 men and 47 women; 18-47 y.o., $M_{age} = 27.3, SD_{age} = 6.5$) 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. Six participants were excluded from the analysis: five of them had an unusually high number of responses below 300 ms at the learning phase and one other participant looped themselves within a single state-response transition by pressing the same button 33 times.

Materials and Procedure

Chatbot 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 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 four probabilistic and six deterministic transitions. The *hard* machine has ten 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) with a two-alternative forced choice question corresponding to the test condition. As a within-subjects 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.

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*

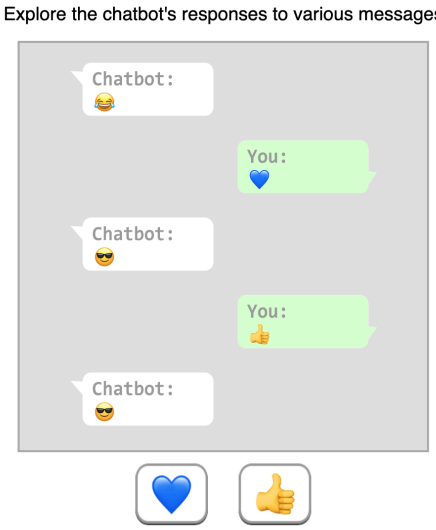

Figure 2: Interaction with the chatbot at the learning ("free exploration") phase.

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

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 $Input_2$.

While our study was designed to determine the extent to which participants could accomplish these tasks, it also has the side-benefit of enabling us to study the learning phase and, in particular, the ways in which participants explored the underlying machine.

Results

Our results are summarized in Table 1 and Fig. 4. The clearest, and strongest, effect is quite simple: after the initial period of free exploration, participants find the prediction and control tasks easier than causal explanation. This effect is strongest when the system is simple (the "easy" machine), and when the test involves fully-visible processes.

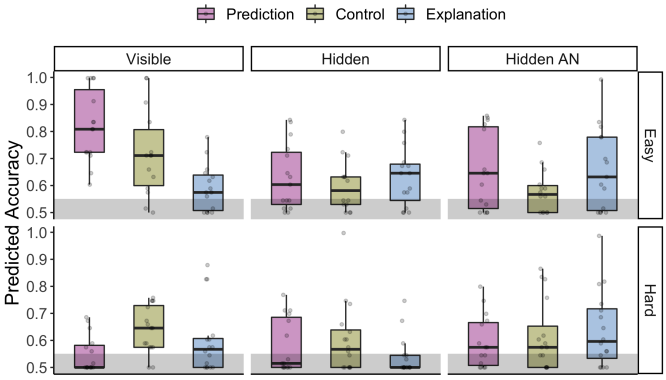

Figure 4: Accuracy values predicted based on participant's mental models for a fixed-difficulty test question. We see a high degree of heterogeneity; a minority of participants achieve good performance on even the most difficult tasks, at high significance. For those who were given the explanation task, high performing participants were best described as using the "alternative neglect" version of casual explanation.

When the problem becomes harder—either because the system to be learned is more complex, or because the participant has to make judgments in the presence of partial information—predicted performance on all three tasks drops significantly. In the harder-visible case, we have some evidence that participants perform better on control than predic-

tion, but effect sizes are small.

Intriguingly, as performance on prediction and control declines, we see one form of explanation—explanation with alternative neglect—rise. For both the easy and hard systems, when the intermediate state is hidden, participants do better on the explanation task, under the assumption that they are using the alternative neglect heuristic.

Fig. 4 shows participant heterogeneity. We have strong evidence that a subset of the population is able to use their mental models to accomplish all three tasks, even in the most difficult cases. In the hidden-hard cases, for example, a fraction of the participants are able to achieve high levels of accuracy; put another way, the poor average performance in Table 1 masks the fact that a minority of participants appear to understand the tasks, and make good use of their mental models.

Examination of this subset of the population also provides additional evidence that explanation with alternative neglect is a better account of how participants undertake causal reasoning. for both the easy and hard systems, many of the participants are well-described as making rigorous use of their mental models to build casual accounts in this fashion.

Task	Easy (Visible)	Easy (Hidden)	Hard (Visible)	Hard (Hidden)
Prediction	$0.82^{0.88}_{0.77}$	$0.63^{0.68}_{0.58}$	$0.55^{0.58}_{0.52}$	$0.58^{0.62}_{0.54}$
Control	$0.72^{0.79}_{0.65}$	$0.60^{0.64}_{0.56}$	$0.64^{0.67}_{0.61}$	$0.60^{0.66}_{0.55}$
Explanation	$0.59^{0.63}_{0.56}$	$0.63^{0.68}_{0.59}$	$0.60^{0.65}_{0.55}$	$0.53^{0.57}_{0.51}$
Expl. with AN	—	$0.66^{0.73}_{0.60}$	—	$0.64^{0.70}_{0.59}$

Table 1: Predicted accuracies for the different conditions, with 5% ranges shown (*i.e.*, one-sided $p < .05$ values). As expected, task performance drops with the more complex machine, and with the "hidden" task. There is some evidence that subjects perform better at control tasks than prediction tasks with the more complex machine. Finally, we have evidence that, when there are hidden variables, performance on explanation—and, in particular, explanation with alternative neglect—exceeds both prediction and control.

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 critical tests 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 learn a task in the presence of feedback, we first present them with a system to explore in an unstructured fashion, allowing them to build a mental model. Only then, in a second phase, do test them on their task performance. Using a simple Bayesian model for performance, we are then able to disentangle the building of a mental model from its deployment in the three tasks in an equal fashion.

Our major result is that, in the simplest cases, people find prediction and control far easier than explanation. 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 would seem to require mastery of causal processes, and thus, some ability to explain—but real-world behavior suggests that control performance is more analogous to prediction.

We have weaker, more preliminary, evidence that suggests control, rather than prediction, persists longer as the tasks become harder, which is in line with the previously discussed studies in a dynamic system control paradigm (Berry & Broadbent, 1984, 1988). Gaining a better sense of how performance on these tasks shifts with difficulty may provide new insight onto foundational debates between prediction-, control-, and explanation-first accounts of human behavior.

Intriguingly, our experiments also provide suggestive evidence for how people do explanation tasks in the more challenging situations. In the hardest problems we study, it is explanation—and, in particular, a particular form of explanation, alternative neglect—that provides the best performance. As performance on prediction and control declines, performance on explanation, if anything, seems to slightly improve. We assume that the hidden intermediate state allows one to reason counterfactually more easily, which is essential in identifying causal relationships.

The major limitation in our work is given by the large heterogeneity in performance. 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: 80% to 90% accuracy, for example, in causal reasoning about complex machines under partial information. As can be seen in Fig. 4, in that subset of high performers, we have good evidence for the weaker trends seen in the population at large: that prediction and control are more reliable in simple systems, but decline quickly as the problems become harder, while explanation, paradoxically, can improve, rather than declines, with challenge.

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*.

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

Clark, A. (2013, June). 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, March). 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, May). 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, February). 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, October). 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, October). 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, July). 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, October). 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, October). 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, October). 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, August). Inferring an Observer's Prediction Strategy in Sequence Learning Experiments. *Entropy*, *22*(8), 896. doi: 10.3390/e22080896

Wojtowicz, Z., & DeDeo, S. (2020, December). 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