# How Predictive Minds Explain and Control Dynamical Systems

**Anonymous Author(s)**
Affiliation
Address
email

## Abstract

We study the relationship between prediction, explanation, and control in artificial "predictive minds"—modeled as Long Short-Term Memory (LSTM) neural networks—that interact with simple dynamical systems. We show how to operationalize key philosophical concepts, and model associated cognitive biases. Our results reveal, in turn, an unexpectedly complex relationship between prediction, explanation, and control. In many cases, "predictive minds" can be better at explanation and control than they are at prediction itself, a result that holds in the presence of heuristics expected under computational resource constraints.

We interact with dynamical systems (DSs) as a basic part of everyday life, and spend a great deal of effort trying to predict, explain, and control the ones we encounter. We often try predict the future of a social or physical DS (e.g., "How would that person react to my words?" "Will the door unlock if I turn the key?"), explain what happened (e.g., "What made them so upset?" "Why did the door unlock"), or attempt to control or guide the system's future (e.g., "How can I make them happy?" "How can I unlock the door?").

These are distinct tasks. Explaining why something went wrong doesn't mean knowing how to fix it. Predicting what will happen doesn't mean being able to explain it. One can control a system without being able to explain how. Unfortunately, the relationship between explanation, prediction, and control remains poorly understood, since the existing methodologies are incapable of measuring them in a comparable manner.

The goal of this paper is twofold. First, we present a novel experimental design that simulates human performance in prediction, explanation, and control, and allows for direct comparison between cognitive theories and the results of behavioral studies. For brevity, we focus on a particularly influential, information-theoretic, account of cognition, the "predictive processing" framework [1, 2, 3], that places prediction—rather than explanation [4, 5, 6], or control [7, 8]—in a privileged position.

Then, we show how differential performance on these tasks can be used as a probe into the nature of cognition itself. What are generic features of performance on the three tasks? How do they change with the underlying cognitive model? What aspects of a system make it easier, or harder, for a predictive mind to learn? By varying the systems that the agent interacts with, we show how to identify universal patterns in the covariance of performance on prediction, explanation, and control.

## 1 Methods

We model a behavioral experiment in which participants observe a probabilistic finite state automaton (see, e.g., Table 4), and are then tested on their ability to predict, explain, and control it. The participants in this work are simulated "predictive minds", equipped with an LSTM module [9]

Submitted to Information-Theoretic Principles in Cognitive Systems Workshop at the 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

provided by the TensorFlow [10] package, which attempts to minimize cross-entropy, i.e., information-theoretic loss. In this preliminary work, we consider a range of different four-state machines with two potential inputs; see Table 4 for two examples. The participant trains by watching the DS evolve on receiving input symbols, and is tested on a variety of tasks that involve processing two steps in time.

For each task, we consider a "visible" version (where the intermediate state is known) and a "hidden" version (where it is not); these are matched so that the surface level complexities of the questions are identical across the tasks. The six tasks are shown schematically in Table 1. To match standard experimental designs in psychology, test items are two-alternative forced choice questions.

| Task | Basic Form | Visible | Hidden |
|---|---|---|---|
| Prediction | "What State will the system end up in?" | $1_a 2_b \underline{?}$ | $1_a \mathrm{X}_b \underline{?}$ |
| Explanation | "Why did the system end up in State 3?" | $1_{\underline{a}} 2_{\underline{b}} 3$ | $1_{\underline{a}} \mathrm{X}_{\underline{b}} 3$ |
| Control | "How would you get the system into State 3?" | $1_{\underline{?}} 2_{\underline{?}} 3$ | $1_{\underline{?}} \mathrm{X}_{\underline{?}} 3$ |

Table 1: Test tasks, and (shorthand) examples of actual questions that might be asked of a simulated reasoner. States encoded as digits, responses as subscript lowercase letters, unknowns as $X$, and queries as "?" (see text).

To model human cognition, we must both specify the normative content of the task—what it would mean to get it right—and describe the way in which a human might actually go about producing an answer. Normative specifications are philosophical: they are accounts of what behavior ought to aim for. Descriptive specifications are scientific: they are theories of how agents answer the question ways that might, or might not, approximate the normative standard.

### 1.1 Normative Operationalization of Prediction, Explanation, and Control

In all cases, the subject is given a partial (two time-step) history of the DS states and input symbols. In the visible **prediction** case, the subject sees (for example) that the DS was in State 1, received Input $a$, transitioned to State 2, and received Input $b$; they are then asked "which state will DS end up in?" The correct answer is $\mathrm{argmax}_i P(i|1_a 2_b)$.

In the visible **explanation** case, the subject sees (for example) that the DS was in State 1, received Input $a$, transitioned to State 2, received Input $b$, and transitioned to State 3. They are then asked "why did the DS end up in State 3?" Let us consider two possible explanations. Option 1: "Because the DS received Input $a$ after State 1, rather than $b$". Option 2: "Because the DS received Input $b$ after State 2, rather than $a$". The correct answer can be evaluated counterfactually; Option 1 is the correct answer if $P(3|1_b 2_b)$ is less than $P(3|1_a 2_a)$; Option 2 is the correct answer in the opposite case.

This operationalizes a natural account of causal reasoning in terms of counterfactuals with either the first input or the second being altered, and everything else fixed as before. The cause of the final state is the action that, were it varied, would make the outcome less likely compared to alternatives. This matches long-standing accounts of explanation as counterfactual causation [11]; importantly, it contrasts with description-length accounts of explanation [12] that rely on a purely Bayesian analysis in terms of material conditionals.

In the visible **control** case, the subject sees (for example) that the DS is in State 1; they are asked to choose a combination of inputs (e.g., $ab$ or $ba$) that will most likely get the DS into State 2, and then State 3. The correct answer is $\mathrm{argmax}_{\{i,j\}} P(3|1_i 2_j)$.

| Task | Visible Form | Hidden Form |
|---|---|---|
| Prediction | $\mathrm{argmax}_i P(i\|1_a 2_b)$ | $\mathrm{argmax}_i \left(\sum_{k=1}^N P(i\|1_a k_b)\right)$ |
| Explanation | $P(3\|1_b 2_b) > P(3\|1_a 2_a)$ | $\sum_{k=1}^N P(3\|1_b k_b) > \sum_{k=1}^N P(3\|1_a k_a)$ |
| Control | $\mathrm{argmax}_{\{i,j\}} P(3\|1_i 2_j)$ | $\mathrm{argmax}_{\{i,j\}} \left(\sum_{k=1}^N P(3\|1_i k_j)\right)$ |

Table 2: Normative answers to the tasks, based on Table 1; here $N$ is the number of intermediate states. In the explanation task, we show the case where the correct answer is "because the first input was $a$"; this example assumes that the DS in question has only two possible input symbols, $a$ and $b$.

The hidden case is identical, but the subject is not told the intermediate state, forcing them to consider different possibilities for the effect of the first action. The normative hidden case is a natural extension of the visible one, however: we simply marginalize over the intermediate state, as shown in Table 2.

### 1.2 Descriptive Operationalization of Prediction, Explanation, and Control

Reasoners are equipped with a "simulation module" (SM), here taken to be an LSTM, that enables them to predict the future state of a DS based on the prior history of both states and user inputs. We assume that the SM gives answers of the form $\hat{P}(i|1_a 2_b)$; *i.e.*, for an arbitrary two-step sequence of both states and inputs, we have an estimate, $\hat{P}$ of the probability that any particular state $i$ will be the result.

If the reasoners are given a perfect SM and unbounded computational abilities, then they will be capable of 100% accuracy, since they can simply substitute $\hat{P}$ for $P$ in Table 2; the mathematical form remains identical. Errors can result in one of two ways.

Firstly, $\hat{P}$ might differ from $P$. The reasoner may have insufficient data, or the SM may be internally inconsistent or miscalibrated. If the DS is a finite state machine, for example, but the SM allows for long-range correlations, the SM may be influenced by spurious coincidences.

Secondly, the reasoner might make systematic errors in how it uses the SM. These errors are at a higher, Marr-algorithmic, level, and include many of the heuristics and biases in the psychological literature. Here, we consider a simple bias, "alternative neglect" (AN), which occurs when the reasoner considers only the most likely alternative for the intermediate case rather than all possible options. This leads to deviations from the normative form, as shown in Table 3.

| Task | Hidden Form | Hidden Form with Alternative Neglect |
|---|---|---|
| Prediction | $\mathrm{argmax}_i \left(\sum_{k=1}^{N} \hat{P}(i\|1_a k_b)\right)$ | $\mathrm{argmax}_i \left(\mathrm{max_k} \hat{P}(i\|1_a k_b)\right)$ |
| Explanation | $\sum_{k=1}^{N} \hat{P}(3\|1_b k_b) > \sum_{k=1}^{N} \hat{P}(3\|1_a k_a)$ | $\mathrm{max}_k \hat{P}(3\|1_b k_b) > \mathrm{max}_k \hat{P}(3\|1_a k_a)$ |
| Control | $\mathrm{argmax}_{\{i,j\}} \left(\sum_{k=1}^{N} \hat{P}(3\|1_i k_j)\right)$ | $\mathrm{argmax}_{\{i,j\}} \left(\mathrm{max}_k \hat{P}(3\|1_i k_j)\right)$ |

Table 3: Descriptive answers for the harder tasks, based on Table 1; compare to Table 2. Here, $\hat{P}$ refers to the output of the simulation module (SM) after training on the prediction task. Errors in the reasoner can come from two paths: a failure to simulate the DS well ($\hat{P}$ deviating from $P$), or a failure to combine $\hat{P}$ in the normative fashion. We consider one failure of the second kind, "alternative neglect"; this only matters for the hidden form of the task.

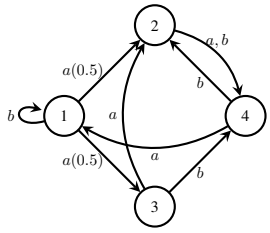
## 2 Results

Table 4 shows the results for two test cases, "DS One" and "DS Two", with a short training set of 32 trials (i.e., single 32-step sequences). Fig. 1 overlays the results for DS One and DS Two with a larger set of 1000 randomly generated systems with similar structure to DS Two; each random system has deterministic transitions except for one state/input case, for which it branches, probabilistically, to three alternatives.

In general, the visible task, which cuts down on possibility space, is usually easier, particularly so for control ($+5$ percentage points easier on average). As expected, alternative neglect heuristic leads to significant deficits in performance; prediction and control decline by $5$ points, and explanation by 7 points, on average, with a broad distribution; Fig. 1(a) shows the explanation case.

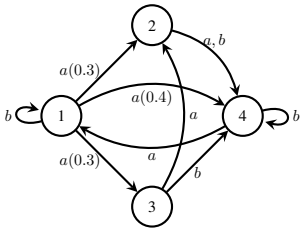
Most importantly, our results provide new insight into the unexpectedly counterintuitive and complex relationships between prediction, explanation and control. Most notably, in a significant fraction of cases—and particularly in the case of the hidden task—our "predictive minds" are actually better at explanation and control than they are at prediction. This is shown in Fig. 1(b) for the hidden form of the task, where we plot the performance on explanation and control relative to prediction. In the hard condition, approximately 58% of the machines are the top right quadrant (48% in hard+AN), where both explanation and control tasks are easier than prediction.

(a) DS One



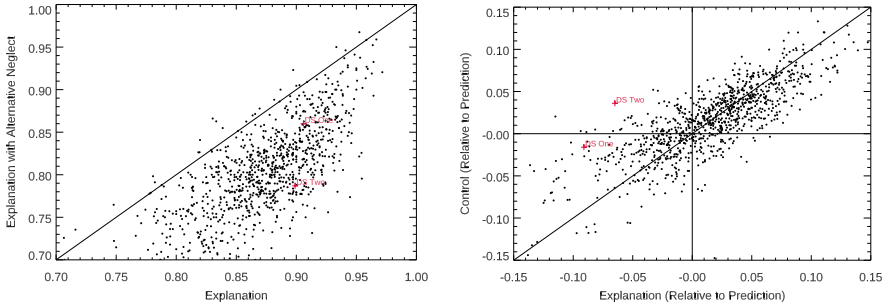

| Task | Visible | Hidden | Hidden+AN |
|---|---|---|---|
| Prediction | 98.7% | 97.3% | 95.3% |
| Explanation | 89.6% | 90.6% | 86.0% |
| Control | 97.1% | 96.4% | 93.2% |

(b) DS Two




| Task | Visible | Hidden | Hidden+AN |
|---|---|---|---|
| Prediction | 94.1% | 91.5% | 89.2% |
| Explanation | 87.6% | 89.9% | 78.7% |
| Control | 97.7% | 94.4% | 89.4% |

Table 4: LSTM accuracy rates for prediction, explanation, and control tasks in two simple machines. The training phase has $N_l = 32$ state-symbol pairs. Standard errors are $< 1\%$.




(a) (b)

Figure 1: Key results: (a) alternative neglect induces systematic deficits in all three tasks, as can be seen here in the case of explanation; (b) it is often easier to control or explain a system than it is to predict one. Dots label randomly-sampled machines; "+" symbols label DS One and DS Two.

## 3 Discussion

Our work begin with the key idea that prediction, explanation, and control ought to be examined simultaneously. A long tradition in dynamic decision-making studies has examined human power to *control* DSs [13, 14, 15]). Prediction and explanation, however, have been largely overlooked. A few works have used DSs to investigate causal inference (e.g., [16]), but did not measure control and prediction. One exception is in implicit learning experiments that included a prediction task in a post-experimental questionnaire, to measure awareness of knowledge obtained about the dynamic system [17, 18]. Consistent with our new results, they found that people can learn to control a dynamic system, but still fail to accurately predict its behaviour.

Our main result is that these effects hold even for "predictive" minds. Agents can find control and explanation easier than prediction—even when their world-models are trained to minimize information-theoretic loss—on prediction tasks. These effects vary from DS to DS, which now provides a guide for laboratory tests that can tease apart the underlying mechanisms associated with each task.

A second result is how normative constraints (Table 2) compare with how a psychologically-realistic subject might actually perform (Table 3). "Alternative neglect" is a common bias in the real world [19]; our work shows how this heuristic leads to characteristic patterns of sub-optimal performance that vary both from DS to DS, and between the tasks (prediction, explanation, and control) themselves.

## References


[1] Andy Clark. Whatever next? predictive brains, situated agents, and the future of cognitive science. 36(3):181–204, 2013. Publisher: Cambridge University Press.

[2] Karl Friston. The free-energy principle: a unified brain theory? 11(2):127–138, 2010.

[3] Jakob Hohwy. *The predictive mind*. Oxford University Press, 2013.

[4] Zachary Horne, Melis Muradoglu, and Andrei Cimpian. Explanation as a cognitive process. 23(3):187–199, 2019.

[5] Tania Lombrozo. The structure and function of explanations. 10(10):464–470, 2006.

[6] Ruth M.J. Byrne. Counterfactual thought. *Annual Review of Psychology*, 67(1):135–157.

[7] David Silver, Satinder Singh, Doina Precup, and Richard S. Sutton. Reward is enough. 299:103535, 2021.

[8] Cleotilde Gonzalez, Javier F. Lerch, and Christian Lebiere. Instance-based learning in dynamic decision making. *Cognitive Science*, 27(4):591–635, 2003. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1207/s15516709cog2704_2.

[9] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. 9(8):1735–1780, 1997. Conference Name: Neural Computation.

[10] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. Tensorflow: Large-scale machine learning on heterogeneous systems, 2015.

[11] Peter Menzies and Helen Beebee. Counterfactual Theories of Causation. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2020 edition, 2020.

[12] Zachary Wojtowicz and Simon DeDeo. From probability to consilience: How explanatory values implement bayesian reasoning. *Trends in Cognitive Sciences*, 24(12):981–993, 2020.

[13] Berndt Brehmer. Dynamic decision making: Human control of complex systems. 81(3):211–241, 1992.

[14] Jared M. Hotaling, Pegah Fakhari, and Jerome R. Busemeyer. Dynamic decision making. In *International Encyclopedia of the Social & Behavioral Sciences*, pages 708–713. Elsevier, 2015.

[15] Cleotilde Gonzalez, Pegah Fakhari, and Jerome Busemeyer. Dynamic Decision Making: Learning Processes and New Research Directions. *Human Factors*, 59(5):713–721, August 2017. Publisher: SAGE Publications Inc.

[16] Mark Steyvers, Joshua B. Tenenbaum, Eric-Jan Wagenmakers, and Ben Blum. Inferring causal networks from observations and interventions. *Cognitive Science*, 27(3):453–489, 2003.

[17] Dianne C. Berry and Donald E. Broadbent. On the Relationship between Task Performance and Associated Verbalizable Knowledge. *The Quarterly Journal of Experimental Psychology Section A*, 36(2):209–231, May 1984.

[18] Dianne C. Berry and Donald E. Broadbent. Interactive tasks and the implicit-explicit distinction. *British Journal of Psychology*, 79(2):251–272, 1988.

[19] Philip M. Fernbach, Adam Darlow, and Steven A. Sloman. Neglect of Alternative Causes in Predictive but Not Diagnostic Reasoning. *Psychological Science*, 21(3):329–336, March 2010. Publisher: SAGE Publications Inc.