A semantic bias for accurate predictive learning in multidimensional environments

Euan Prentis (eprentis@uchicago.edu)
Department of Psychology, University of Chicago, Chicago, Illinois, USA

Akram Bakkour (bakkour@uchicago.edu)
Department of Psychology, University of Chicago, Chicago, Illinois, USA

Abstract

To make accurate inferences about our multidimensional world, humans must distinguish observations of causal processes from spurious associations. We investigated the role of inductive biases in shaping memory around causal information, specifically testing for a semantic bias that leverages existing semantic structure to direct learning. Participants completed a predictive learning task in which both causal and spurious associations were observed. Results showed that spurious inferences were suppressed when the causal associations were defined within semantic categories, indicating that a semantic bias directed learning. Simulations of a feature-based successor features model further demonstrated that this bias should have a more dramatic benefit in more naturalistic environments, with high-dimensional states and deep causal processes. In all, this work demonstrates that inductive biases that act on multidimensional transition dynamics may be essential for learning in our complex world.

Keywords: Predictive Learning; Decision Making; inductive Biases; Reinforcement Learning

Introduction

To make good decisions, humans must predict how our actions affect future events. This ability is supported by predictive memory representations that encode the statistical relationships between events (Momennejad, 2020; Stachenfeld et al., 2017). However, learning useful representations is not straightforward given the complexity of real-world experience. Events comprise numerous features, whose many causal interactions determine how experience unfolds. These interactions may occur in parallel, creating ambiguity about which observations reflect causal versus spurious associations (Liljeholm, 2020). For example, say causal processes A1→A2 and B1→B2 co-occur ({A1, B1} → {A2, B2}; Figure 1), an observer will incidentally witness spurious transitions A1→B2 and B1→A2. These spurious associations may distort the learned predictive representation, leading to noisy inference about the outcomes of events comprising A1 or B1.

When a set of features *frequently* co-occur, these distortive effects may do more than inject noise – they may tune learning to specific contexts. Spurious transitions in effect act as links between disparate causal processes. These links are reinforced as causal processes co-occur, binding features into common representations that reflect the context defined by the feature conjunctions (e.g., {A1, B1} → {A2, B2}) rather than the independent causal processes (e.g., A1→A2, B1→B2). Thus, multidimensional environments are inherently prone to noisy, warped representation.

Inductive biases can direct learning to stable properties across contexts (Goyal & Bengio, 2022; Kemp & Tenenbaum, 2009). The current work explores how inductive biases may thus direct predictive learning to context-independent causal processes. Specifically, we test for a *semantic bias*, whereby learning is molded towards the existing structure of semantic knowledge.

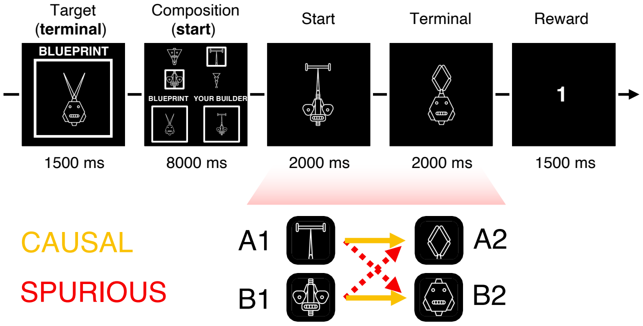

Figure 1: Trial procedure. START→TERMINAL state transitions arise from feature-based causal transitions.

Semantic bias shapes predictive learning

One hundred participants completed a feature-based predictive learning task in which they made compositional inferences to earn reward (Figure 1). Stimuli were robot images, each comprising two features drawn from four semantic categories (heads, arms, bodies, antennas). Causal associations were defined between “start” and “terminal” features. On each trial, participants saw a “target” terminal item before a set of four start features. Their goal was to select two start features to compose an item that would produce the target. After choosing, the terminal item produced by the composition was shown, and a reward was paid based on the number of features that matched between the target and produced terminal items.

To test for a semantic bias, participants were assigned to one of two between-subjects conditions. In the semantic congruent condition, causal transitions were defined within category (e.g., head1→head2). In the semantic incongruent condition, causal transitions were defined between categories (e.g., head1→arm2). A semantic bias could only apply in the congruent condition, potentially directing learning to causal over spurious transitions. To estimate each participant’s extent of causal and spurious learning, we fit a multinomial logistic model that quantified the influence of observed causal and spurious transitions on choice. Consistent with a semantic bias, causal versus spurious transitions influenced choice more in the semantic congruent condition (M = 0.875, SD = 0.358, 95% HDI = [0.187, 1.588]; Figure 2A).

Next, we tested whether this bias reduced learning specificity. We had half the robots occur twice as often as the others, with the expectation that high specificity learning should tune to frequent trials, diminishing reward earnings on infrequent trials. Supporting our hypothesis, reward earnings were more sensitive to trial frequency in the incongruent condition (M = 0.141, SD = 0.034, 95% HDI = [0.077, 0.208]; Figure 2B).

Modelling learning in complex contexts

While our task captured key aspects of multidimensional experience, it remained far simpler than real-world environments, which comprise higher-dimensional events and deeper causal processes that unfold over extended durations. We therefore sought to understand how the observed semantic bias may affect learning in more naturalistic environments.

Individual differences in predictive learning were characterized by fitting a feature-based successor features model (FBSF) to each participant's data. The standard SF model is a predictive representation that encodes the quantity of future features expected to be encountered from a state (S→F'; (Carvalho et al., 2024; Dayan, 1993). To model learning fully at the level of features, FBSF instead encodes expectations for each *feature* (F→F'). Difference in bias were captured by a free parameter (bias = [0, 1]) that dictated the suppression of spurious information during learning.

To verify that FBSF suitably captured learning, we compared it to three alternative models: (1) CBSF – learned state transitions (S→S'); (2) CBSF sampler – CBSF with a retroactive integration mechanism for flexible inference; and (3) null – random choice. AIC-based model selection found that most participants not best fit by the null model were best fit by FBSF (Figure 2C). Moreover, verifying that FBSF in isolation explained behavior, the bias fit was associated with greater causal versus spurious transition influence (M = 2.692, SD = 0.691, 95% HDI = [1.280, 4.012]; Figure 2D) and reduced frequency sensitivity (M = -0.239, SD = 0.059, 95% HDI = [-0.349, -0.120]; Figure 2E).

Finally, to extrapolate behavior to more naturalistic settings, we simulated the best fitting FBSF models in task environments varying in dimensionality (1, 2, or 4 features per state) and causal depth (1-, 2-, 3-, or 4-step). Models were trained to convergence, and then tested on trials with no feedback. Agents earned less reward in environments with greater dimensionality and depth (Figure 2F). However, high-bias agents exhibited a striking reduction in this impairment, highlighting the crucial role such inductive biases might have for predictive learning in real-world environments.

Conclusion

These results provide evidence for a semantic bias that molds predictive representations based on existing semantic knowledge. This bias suppressed spurious information, improving the accuracy of predictive inference, and limiting learning specificity. Formal modeling further demonstrated that these benefits scale with the dimensionality and causal depth of the environment. Thus, inductive biases that act on multidimensional transition dynamics may be crucial for effective predictive learning in complex everyday life.

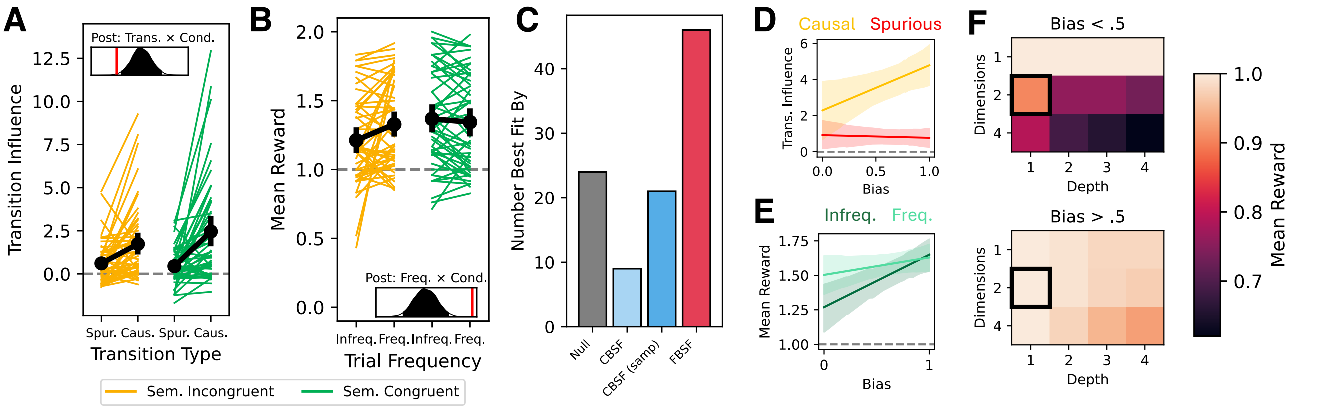

Figure 2: Results. (**A**) Influence of causal versus spurious transitions on choice. (**B**) Mean reward earnings by trial frequency and condition. (**C**) Model fits. (**D**) Transition influence on choice by bias fit. (**E**) Sensitivity of reward earning to trial frequency by bias fit. (**F**) Simulated FBSF reward earnings by environment depth and dimensionality.

References

Carvalho, W., Tomov, M. S., de Cothi, W., Barry, C., & Gershman, S. J. (2024). *Predictive representations: Building blocks of intelligence* (arXiv:2402.06590). arXiv. http://arxiv.org/abs/2402.06590

Dayan, P. (1993). *Improving Generalisation for Temporal Difference Learning: The Successor Representation*. 14.

Goyal, A., & Bengio, Y. (2022). Inductive biases for deep learning of higher-level cognition. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *478*(2266), 20210068. https://doi.org/10.1098/rspa.2021.0068

Kemp, C., & Tenenbaum, J. B. (2009). Structured statistical models of inductive reasoning. *Psychological Review*, *116*(1), 20–58. https://doi.org/10.1037/a0014282

Liljeholm, M. (2020). Neural Correlates of Causal Confounding. *Journal of Cognitive Neuroscience*, *32*(2), 301–314. https://doi.org/10.1162/jocn_a_01479

Momennejad, I. (2020). Learning Structures: Predictive Representations, Replay, and Generalization. *Current Opinion in Behavioral Sciences*, *32*, 155–166. https://doi.org/10.1016/j.cobeha.2020.02.017

Stachenfeld, K. L., Botvinick, M. M., & Gershman, S. J. (2017). The hippocampus as a predictive map. *Nature Neuroscience*, *20*(11), 1643–1653. https://doi.org/10.1038/nn.4650