Controllability Governs the Balance Between Pavlovian and Instrumental Action Selection

Hayley M. Dorfman & Samuel J. Gershman
Department of Psychology and Center for Brain Science, Harvard University

Keywords
Pavlovian, instrumental, agency, Bayesian inference, reinforcement learning

Abstract

A Pavlovian bias to approach reward-predictive cues and avoid punishment-predictive cues can conflict with instrumentally-optimal actions. While most previous work has assumed that this bias is a fixed individual trait, we argue that it can vary within an individual. In particular, we propose that the brain arbitrates between Pavlovian and instrumental control by inferring which is a better predictor of reward. The instrumental predictor is more flexible; it can learn values that depend on both stimuli and actions, whereas the Pavlovian predictor learns values that depend only on stimuli. The cost of this flexibility is error due to overfitting, since a more flexible predictor can more easily fit randomness in the data. The arbitration theory predicts that the Pavlovian predictor will be favored when rewards are relatively uncontrollable, because the additional flexibility of the instrumental predictor is not useful. Consistent with this hypothesis, the Pavlovian approach bias is stronger under low control compared to high control contexts.

Introduction

Pavlovian processes promote approach towards reward-predictive stimuli and avoidance of punishment-predictive stimuli (Wasserman, Franklin, & Hearst, 1974), even when they produce maladaptive behavior (K. Breland & Breland, 1961). For example, Hershberger (Hershberger, 1986) famously demonstrated that newborn chicks struggled to learn that they should walk away from a cup of food in order to obtain it. The chicks could not suppress their Pavlovian tendency to move toward the cup, which was rigged to move farther away as the chicks approached. Another example of Pavlovian misbehavior comes from studies of autoshaping, in which animals interact with a reward-predictive cue (e.g., pigeons will peck a keylight that precedes pellet delivery) despite the fact that these behaviors do not affect the reward outcome. If an omission contingency is then introduced, such that expression of these behaviors causes the reward to be withheld, animals will sometimes persist in performing the maladaptive behavior, a phenomenon known as "negative automaintenance" (D. R. Williams & Williams, 1969). Humans also exhibit Pavlovian misbehavior in Go/No-Go tasks, erroneously acting in response to reward-predictive stimuli when they should withhold action, and erroneously withholding action in response to punishment-predictive stimuli when they should act (Guitart-Masip, Duzel, Dolan, & Dayan, 2014; Guitart-Masip, Huys, et al., 2012b).

The idea that instrumental and Pavlovian processes coexist and compete for control of behavior has been a long-standing fixture of associative learning theory (Miller & Konorski,

1928; Mowrer, 1947; Rescorla & Solomon, 1967), and more recently has been formalized within the framework of modern reinforcement learning theories (Dayan, Niv, Seymour, & Daw, 2006). These theories have typically assumed that instrumental and Pavlovian processes each provide action values, which are then linearly combined to produce composite action values that control behavior. A weighting parameter determines the degree of Pavlovian influence, and this parameter is fit to each participant in the experimental data set. The purpose of the present paper is to revisit the assumption that the weighting parameter is fixed within an individual, and instead argue that the weighting parameter is determined endogenously by an arbitration process, much like an influential proposal for the arbitration between model-based and model-free reinforcement learning strategies (Daw, Niv, & Dayan, 2005).

Our theory of arbitration is based on the idea that Pavlovian and instrumental processes can be understood as constituting different predictive models of reward (we will use the terms 'predictor' and 'model' interchangeably, except where we distinguish the brain's internal models of the environment from our models of the brain). The instrumental predictor learns reward expectations as a function of both stimuli and actions, whereas the Pavlovian predictor learns reward expectations as a function only of stimuli. Thus, the instrumental predictor is strictly more complex than the Pavlovian predictor: it can capture any pattern that the Pavlovian predictor can capture, as well as patterns that the Pavlovian predictor cannot capture. The cost of this flexibility is that the instrumental predictor can also overfit on a finite data set, which means that it will generalize poorly due to fitting noise. The basic problem of arbitration is thus to negotiate a balance between capturing the patterns in the data (favoring the more complex instrumental predictor) and avoiding overfitting (favoring the less complex Pavlovian predictor).

Bayesian model averaging elegantly resolves this problem by weighting each predictor's output by the posterior probability of the predictor given the data. The posterior will tend to favor predictors of intermediate complexity, due to what is known as *Bayesian Occam's razor* (MacKay, 2003). We can think of each predictive model as 'betting' on observing particular data sets (Fig. 1, left). Simple models concentrate their bets on a relatively small number of data sets, whereas complex models distribute their bets across a larger number of data sets. If a simple model accurately predicts a particular data set, it is 'rewarded' more than a complex model, because it bet more on that data set. If the model is too simple (i.e., its bets are too narrowly concentrated), it will fail to predict the observed data.

Another perspective on the same idea comes from the bias-variance trade-off (Geman, Bienenstock, computation, 1992, 1992; Gigerenzer & Brighton, 2009; Glaze, Filipowicz, Kable, Balasubramanian, & Gold, 2018). Any predictor's generalization error (i.e., how poorly it predicts new data after learning from a finite amount of training data) can be decomposed into the sum of three components: squared bias, variance, and irreducible error. Bias is the systematic error incurred by adopting an overly simple model that cannot adequately capture the underlying regularities in the data. Variance is the random error incurred by adopting an overly complex model, which causes the model to overfit random noise in the training data. The irreducible error arises from the inherent stochasticity of the data-generating process, which is independent of the predictor. Bias can be reduced by increasing model complexity, but at the cost of increasing variance. Optimal generalization error is achieved at an intermediate level of complexity where the sum of squared bias and variance (i.e., the reducible error) is minimal (Fig. 1, right). The bias-variance trade-off is closely connected to the Bayesian model averaging perspective, because predictive models with higher posterior probability will tend to have lower generalization error (Germain, Bach, Lacoste, & Lacoste-Julien, 2016).

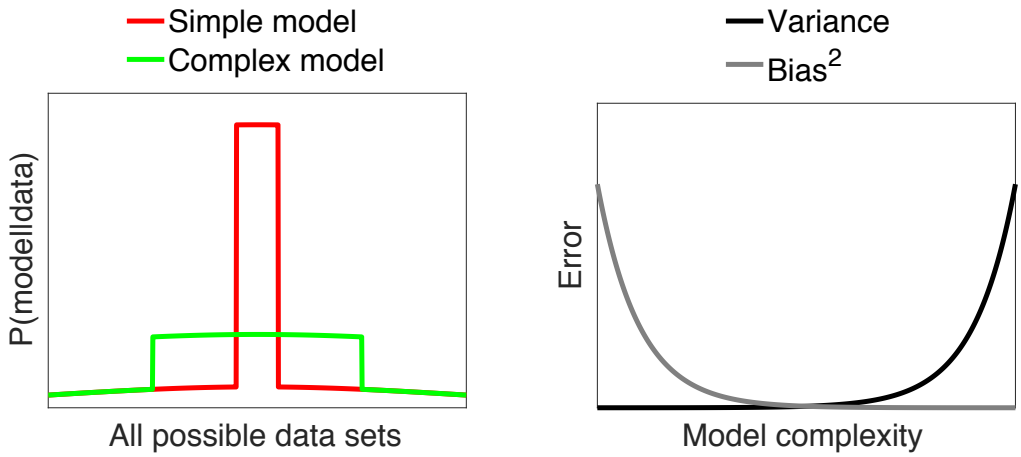


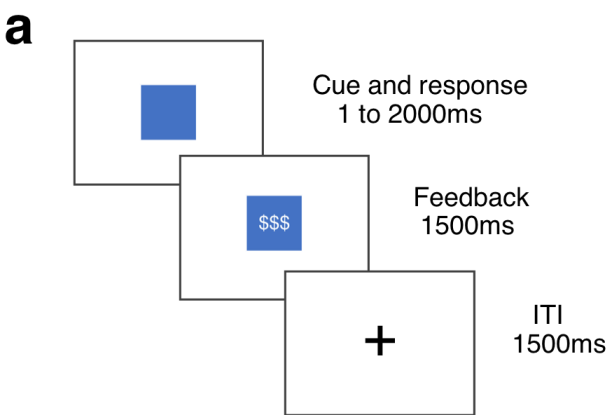
Fig. 1: Two perspective on model complexity. (Left) Bayesian Occam's razor. Complex distribute their probability mass across many different data sets, and thus get less credit for observing any particular data set, whereas simple models concentrate their probability mass on a small number of data sets, and thus get relatively more credit when those data sets are observed. (Right) As model complexity increases, generalization error due to bias decreases, while generalization due to variance increases.

Applying these ideas to arbitration between Pavlovian and instrumental control, a key determinant of the optimal model complexity is *controllability of reward* (Huys & Dayan, 2009; Moscarello & Hartley, 2017). If rewards are uncontrollable (actions do not affect reward rate), then the simpler Pavlovian predictor will be favored by the posterior, because the additional complexity of the instrumental predictor is not justified relative to the penalty imposed by the Bayesian Occam's razor. Only when rewards are sufficiently controllable, or once sufficient data have been observed, will the instrumental predictor be favored (asymptotically, the instrumental predictor will always be favored, because the risk of overfitting noise disappears as the data set becomes large).

We test the predictions of the Bayesian arbitration model by manipulating reward controllability in two Go/No-Go experiments, using the Pavlovian go bias observed in previous experiments (Cavanagh, Eisenberg, Guitart-Masip, Huys, & Frank, 2013; Guitart-Masip, Chowdhury, et al., 2012a) as an index of Pavlovian control. As a complementary window into the arbitration process, we also explore how controllability affects the bias-variance trade-off.

Method

We describe the two experiments together because they are very similar in structure (Fig. 2). Experiment 1 manipulated reward controllability between participants, whereas Experiment 2 manipulated it within participants.



b

	Go	No-Go
Go-to-Win	0.75	0.25
No-Go-to-Win	0.25	0.75
Low Control	0.5	0.5
High Control	0.8	0.2

Fig. 2: Behavioral task details. (a) Participants viewed a colored shape cue (for up to 2s) and had to decide whether to press the space bar (Go) or refrain from pressing the space bar (No-Go). They then received feedback (1.5s) denoted by dollar signs (reward) or a rectangular cue (neutral). Participants were instructed that they would receive a small amount of real bonus money for each rewarded outcome, and no bonus money for each neutral outcome. Feedback was followed by an inter-trial-interval (ITI, 1.5s). (b) Reward contingencies for each trial type (Go-to-Win; No-Go-to-Win) by action type (Go; No-Go) by task condition (Low Control; High Control). Task condition was manipulated either between (Experiment 1) or within (Experiment 2) participants.

Participants

We recruited two independent samples of adults from Amazon Mechanical Turk (Experiment 1: $N = 189$, Experiment 2: $N = 212$). The sample sizes were chosen in order to exceed sample sizes from previous, similar work (Guitart-Masip, Huys, et al., 2012b; Guitart-Masip, et al., 2014), Participants for Experiment 2 were recruited from an existing pool of Amazon Mechanical Turk workers. These workers have completed previous experiments for our lab and expressed interest in being re-contacted for additional study opportunities. Participants were excluded for inaccuracy. Specifically, if participants made the incorrect action (either a button press for a No-Go trial, or the absence of a button press for a Go trial) for ≥ 50% of all trials, they were excluded from analyses. This left a total of 98 accurate participants for Experiment 1 and 183 accurate participants for Experiment 2.

Procedure

Participants completed a modified Go/No-Go paradigm, inspired by previous work (Cavanagh et al., 2013; Guitart-Masip, Huys, et al., 2012b). Participants viewed a single colored square on each trial and had to learn the appropriate response for each square. There was a different correct response and reward probability combination for each shape: One square was a Go stimulus, where a spacebar press was rewarded 75% of the time, one square was a No-Go stimulus, where the absence of a button press was rewarded 75% of the time, and the third square was a *Decoy* stimulus, where a spacebar press was rewarded with a particular probability, which was manipulated based on experimental condition. In the *Low control* (LC) condition, the Decoy was rewarded 50% of the time, and in the *High control* (HC) condition – the Decoy was rewarded 80% of the time. Our task differed from previous Go/No-Go tasks in that it did not include any punishment conditions. Rewarded outcomes were represented with dollar signs, and unrewarded outcomes were represented with a neutral (white rectangle) cue. Participants were told that they would receive a small amount of real bonus money for each reward outcome, and their total bonus was summed and disclosed at the end of the experiment.

In Experiment 1, participants were randomly assigned to one decoy condition (LC or HC), so that each participant was exposed to three different stimuli (Go-to-Win, No-Go-to-Win, and either LC or HC). The experiment consisted of 120 trials, 40 trials for each type of stimulus, randomly interleaved. In Experiment 2, each participant experienced both decoy conditions in separate blocks. The experiment consisted of 240 trials, 120 for each block, with 40 trials for each stimulus within a block.

Computational model

On each trial of the task, the participant must take an action (a) in response to a stimulus (s) in order to receive a reward (r). The problem facing the participant is to determine whether they are acting in an environment where outcomes are controllable (instrumental) or uncontrollable (Pavlovian).

Each model has a set of parameters θ that must be learned. The parameters for the uncontrollable model are indexed only by the stimulus (θ_s), whereas the parameters for the controllable model are indexed by both the stimulus and action (θ_{sa}). We will walk through the learning equations for the uncontrollable model, but the idea is essentially the same for the controllable model.

The posterior over parameters given data $\mathcal{D}$ (the history of stimuli, actions and rewards) and environment $m \in \{controllable,\ uncontrollable\}$ is stipulated by Bayes' rule:

$$P(\theta|\mathcal{D},m) \propto P(\mathcal{D}|\theta,m)P(\theta|m)$$

where $P(\mathcal{D}|\theta,m)$ is the likelihood of the data given hypothetical parameter values θ, and $P(\theta|m)$ is the prior probability of those parameter values. In the context of our task, where rewards are binary, $\theta_s = \mathbb{E}[r|s]$ corresponds to the mean of a stimulus-specific Bernoulli distribution. When $P(\theta_s)$ is a $Beta(\theta_0 \frac{\eta_0}{2}, (1-\theta_0)\frac{\eta_0}{2})$ distribution, the posterior mean $\hat{\theta}_s$ (which is also the posterior predictive mean for reward) is initialized to θ_0 and updated according to:

$$\Delta\hat{\theta}_s = \eta_s^{-1}\delta$$

where δ is the reward prediction error ($r - \hat{\theta}_s$), and η_s^{-1} is the learning rate with counter η_s initialized to η_0 and incremented by 1 every time stimulus s is encountered (in the controllable model, η is indexed by both s and a). Intuitively, θ_0 corresponds to the prior mean (the reward expectation before any observations), and η_0 corresponds to the prior confidence (how much deviation from the prior mean the agent expects).

Because the true environment is unknown, it must be inferred, which can be done using another application of Bayes' rule:

$$P(m|\mathcal{D}) \propto P(\mathcal{D}|m)P(m)$$

where

$$P(\mathcal{D}|m) = \int P(\mathcal{D}|\theta, m)P(\theta)d\theta$$

is the marginal likelihood. The posterior can be updated in closed form. For clarity we adopt a log-odds convention, with the prior log-odds given by:

$$L_0 = \log \frac{P(uncontrollable)}{P(controllable)}$$

The posterior log odds are initialized to L_0 and updated according to:

$$\Delta L = r \log \frac{\hat{\theta}_s}{\hat{\theta}_{sa}} + (1 - r) \log \frac{1 - \hat{\theta}_s}{1 - \hat{\theta}_{sa}}$$

Finally, we need to specify how each model maps reward predictions onto action values. For the instrumental model, we assume that action values simply correspond to the expected reward for a particular state-action pair: $V_I(s, a) = \hat{\theta}_{sa}$. For the Pavlovian model, we assume that the action value is equal to $V_P(s, a) = 0$ for a = No-Go and $V_P(s, a) = \hat{\theta}_s$ for a = Go. This assumption follows from the influential idea that Pavlovian reward expectations invigorate action (Guitart-Masip et al., 2014). To combine the two action values into a single integrated value for action selection, we weight each model's value by its corresponding posterior probability:

$$V(s, a) = wV_P(s, a) + (1 - w)V_I(s, a),$$

where

$$w = P(m = uncontrollable|\mathcal{D}) = \frac{1}{1 + e^{-L}}$$

is the posterior probability of the uncontrollable environment.

To allow for stochasticity of behavior, we model the agent's action selection according to a softmax, where β is an inverse temperature parameter controlling the level of choice stochasticity:

$$P(a|s) = \frac{\exp\left[\beta V(s, a)\right]}{\sum_{a'} \exp\left[\beta V(s, a')\right]}$$

The model outlined above, which we will refer to as the *adaptive model*, updates the weighting parameter from trial-to-trial based on the relative predictive accuracy between the two controllers. We also fit a comparison model, which instead fits the weighting term as a free parameter. We refer to this comparison model as the *fixed model*. The models share the same

underlying information processing architecture (Fig. 3) but differ in whether w is set exogenously (in the case of the fixed model) or endogenously (in the case of the adaptive model).

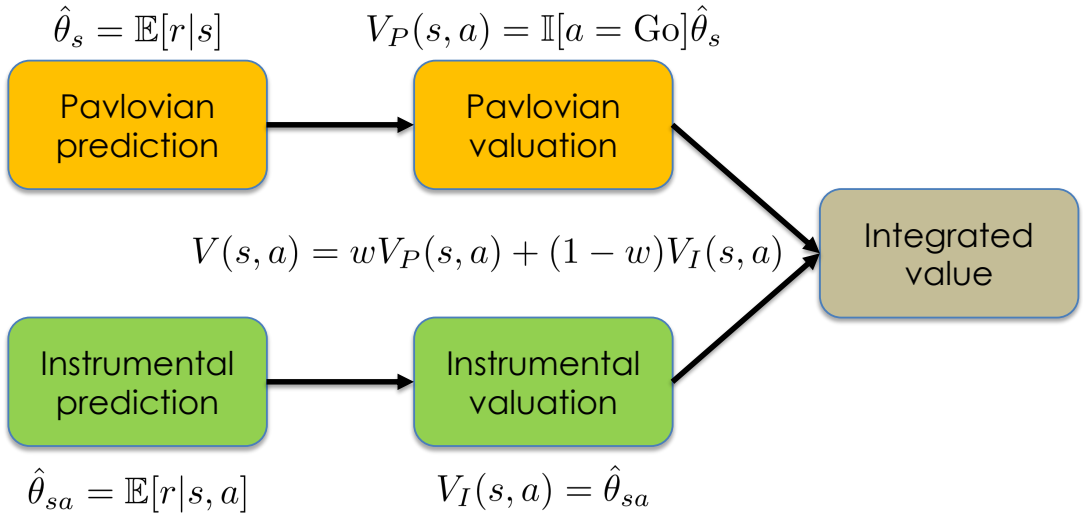


Fig. 3: Information processing architecture. Pavlovian and instrumental prediction and valuation combine into a single value. This integrated value includes a weighting parameter (w) that represents the evidence for the uncontrollable environment (i.e., in favor of the Pavlovian predictor).

We fit each model's free parameters using maximum likelihood estimation. The adaptive model had five free parameters: the inverse temperature β, and the parameters of the prior (θ_0, η_0) for each environment (High or Low Control). We also considered a model in which L_0 was fit as a free parameter, but model comparison indicated that fixing $L_0 = 0.5$ had greater support in our data sets. The fixed model had six free parameters: the same five as the adaptive model, plus the weighting parameter w.

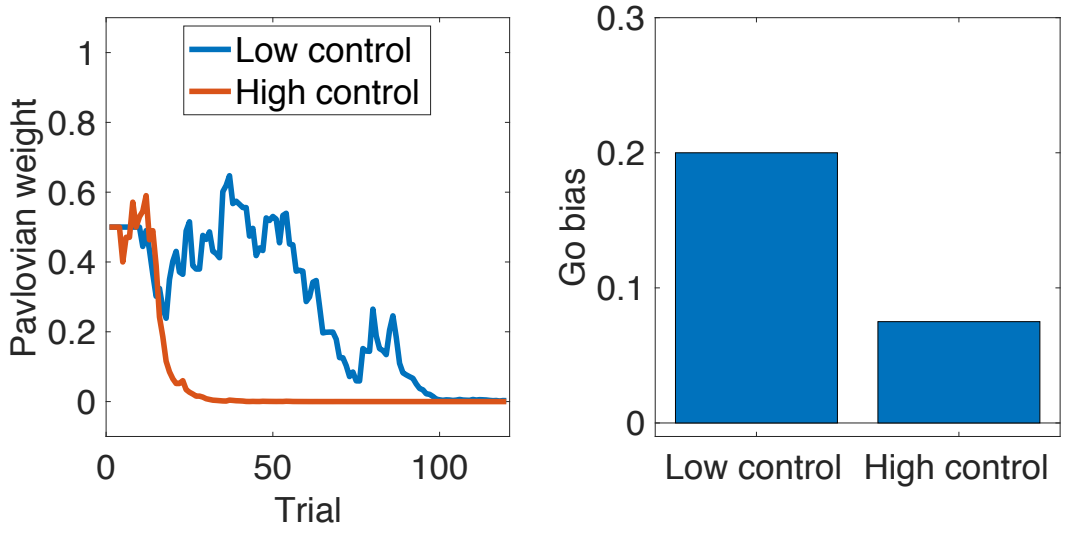


Fig. 4: Model Simulations. Our model demonstrates a greater reliance on the Pavlovian system in the Low Control condition compared to the High Control condition (a) Across all trial types, the Pavlovian weight derived from the model is greater for the Low Control condition. (b) The model can also account for a greater *Go bias* in the Low Control condition. The Go-bias is the difference in accuracy between Go and No-Go trials.

Bias-variance analysis

To assess how controllability affects the bias-variance trade-off, we calculated these quantities for each participant as follows:

$$bias = \sum_{n=1}^{N} \mathbb{I}[a_n = Go] - \mathbb{I}[a_n^* = Go]$$

$$variance = \sum_{n=1}^{N} (\mathbb{I}[a_n = Go] - \bar{a}_n)^2$$

where a_n is the chosen action on trial n, a_n^* is the optimal action, $\bar{a}_n = \frac{1}{N}\sum_{n=1}^{N} \mathbb{I}[a_n = Go]$, and $\mathbb{I}[\cdot] = 1$ when its argument is true, and 0 otherwise.

Results

To investigate the extent to which participants relied on Pavlovian control, we measured their *Go bias*, defined as the accuracy difference between Go and No-Go trials. Under purely instrumental control, the Go bias should be 0, hence values greater then 0 indicate the influence of Pavlovian control. Figure 4 shows simulations of the adaptive model under high and low control conditions, demonstrating the prediction that low control should produce a higher Pavlovian weight (*w*) on average, which will in turn cause a stronger Go bias.

Consistent with the model simulation, participants across both experiments showed an increased go-bias in the LC condition compared to the HC condition (Fig. 5). We used non-parametric tests to test for differences due to the non-normality of the data, as determined by a Lilliefors test. Specifically, a Mann-Whitney U-test revealed a significant difference between the Go bias in the HC and LC condition for Experiment 1 ($p < 0.001$), and a Wilcoxon signed rank test revealed significant differences between conditions for Experiment 2 ($p < 0.05$). The effect appears to be smaller in the within-participant design (Experiment 2), possibly due to cross-talk between conditions.

The adaptive model provided a quantitatively superior account relative to the fixed model, as assessed by random effects Bayesian model comparison (Stephan, Penny, Daunizeau, Moran, & Friston, 2009). Specifically, we calculated the protected exceedance probability (*PXP*), the probability that a particular model is more frequent in the population than all other models under consideration, taking into account the possibility that some differences in model evidence are due to chance. For both experiments, the PXP favoring the adaptive model was >0.99.

To verify the quantitative accuracy of the adaptive model, we plotted the Go bias as a function of weight quantile (Fig. 5), finding a close fit between model and data (for both experiments, the signed rank test comparing the Go bias for the lowest and highest quantiles was significant at $p < 0.001$). Importantly, the quantiles were computed within participants, demonstrating that the model can capture variations in Pavlovian control over the course of a single experimental session.

The timeseries of weights generated by the adaptive model is, on average, strongly correlated with the parameter estimates obtained from fitting the fixed model ($r = 0.88$, $p < 0.0001$). This demonstrates that the adaptive model's average behavior produces behavior similar to that predicted by earlier models using fixed weights (e.g., Guitart-Masip et al., 2012; Cavanagh et al., 2013), but with the weight determined endogenously rather than fit as a free parameter.

We also tested the prediction that the Go bias should diminish over the course of training, and eventually disappear, as can be seen in the simulations (Fig. 4). Consistent with this prediction, the Go bias in Experiment 1 was greater for the first 40 trials compared to the last 40 trials ($p < 0.001$, signed rank test; Fig. 6). The prediction is harder to test in Experiment 2, where the early trials of one condition occur after the late trials of the other condition.

Finally, we examined the effect of controllability on the bias-variance trade-off (Fig. 7). Because controllability favors the more complex instrumental model, we hypothesized that the HC condition would produce lower bias and higher variance (note that this bias should not be confused with the Pavlovian Go bias). This prediction was confirmed in both Experiment 1 (Mann-Whitney U-tests for bias, $p < 0.001$, and variance: $p < 0.001$) and Experiment 2 (Signed rank tests for bias, $p < 0.05$, and variance: $p < 0.001$).

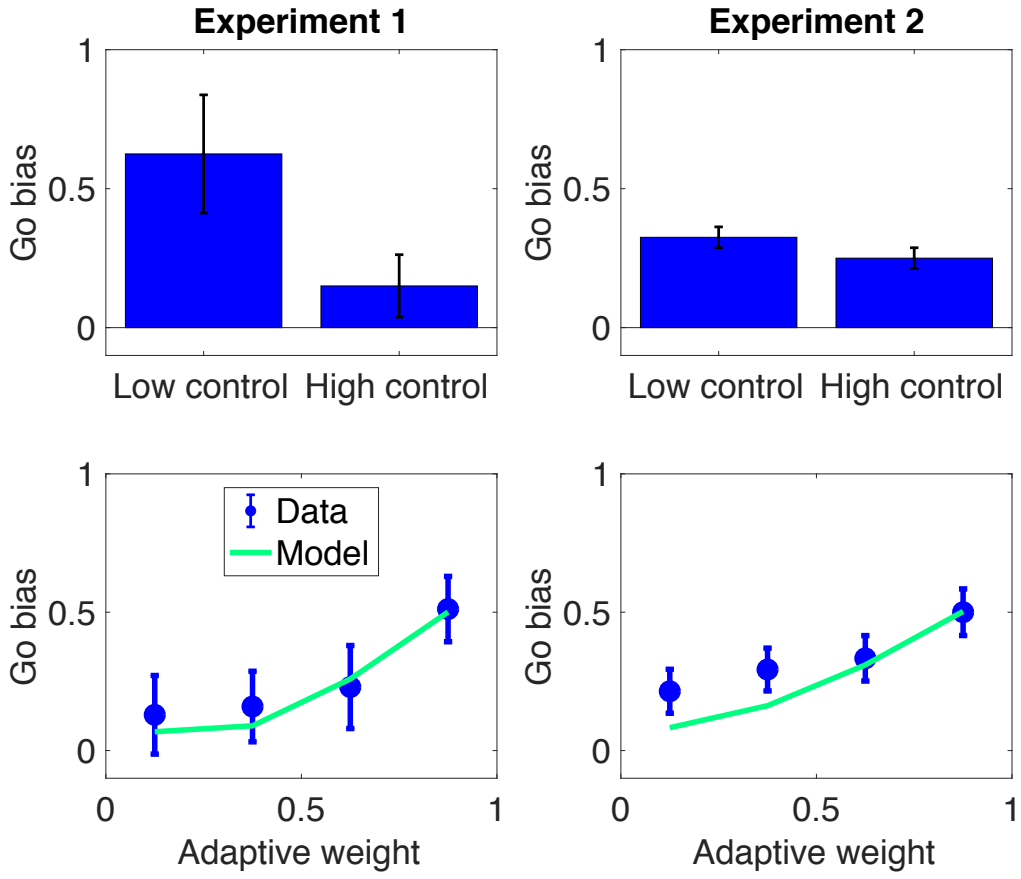


Fig. 5: (Top) Go bias for low and high control conditions in Experiment 1 (left) and Experiment 2 (right). (Bottom) The adaptive model captures within-participant variability in Go bias, plotted as a function of Pavlovian weight (w)

quantile for Experiment 1 (left) and Experiment 2 (right). Error bars show bootstrapped 95% confidence intervals around the median.

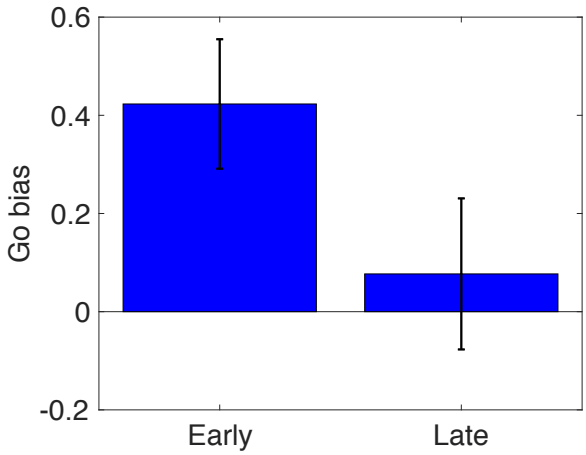


Fig. 6: Go bias in Experiment 1 is larger on the first 40 trials compared to the last 40 trials. Error bars show bootstrapped 95% confidence intervals around the median.

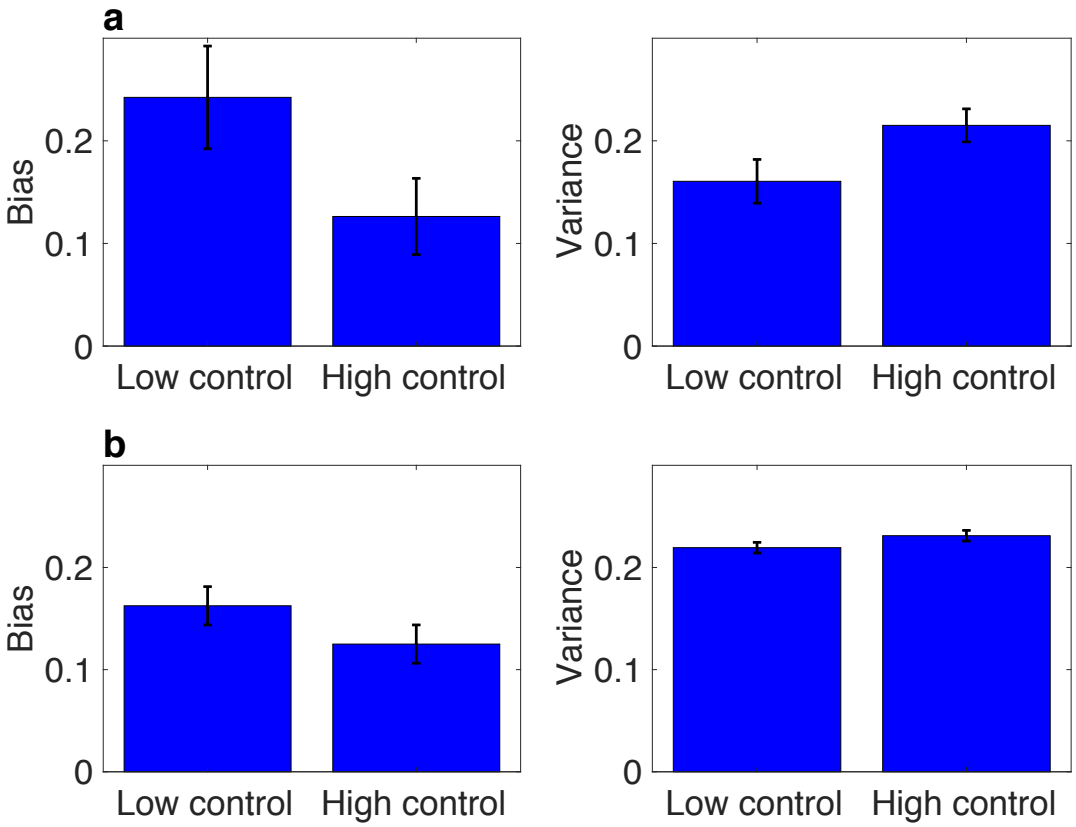


Fig. 7: Bias and variance of choice behavior for Experiment 1 (top) and Experiment 2 (bottom). Error bars show bootstrapped 95% confidence intervals around the median.

Discussion

Taken together, our experimental data provide support for a Bayesian model averaging model of Pavlovian-instrumental arbitration. Our key finding was that the Pavlovian Go bias was stronger under conditions of low reward controllability, consistent with the model's prediction. Analyses in terms of the bias-variance trade-off supported the same conclusion: low controllability favors the simpler Pavlovian predictor, leading to high bias and low variance.

The idea that Pavlovian-instrumental interactions are governed by probabilistic inference joins a number of related ideas in the theories of reinforcement learning. Most relevantly, Daw et al. (2005) suggested that arbitration between model-based and model-free control was determined by Bayesian arbitration, but they did not address Pavlovian-instrumental interactions. A number of earlier theories argued that certain reinforcement learning behaviors could be understood as arising from a model comparison process (Courville, Daw, & Touretzky, 2006; Gershman, 2017; Gershman, Blei, & Niv, 2010; Tomov, Dorfman, & Gershman, 2018). However, to our knowledge, ours is the first account that directly addresses Pavlovian-instrumental interactions in terms of model comparison/averaging.

Our results suggest several directions for future work. First, we have only studied the dynamics of the Pavlovian go bias for rewards; earlier work (e.g., Guitart-Masip et al., 2012) suggests that we should find a symmetric pattern for punishments, with a stronger No-Go bias under low controllability. Second, neuroimaging could be used to identify the neural correlates of arbitration. If our account is correct, we would expect to see a signal in the brain that encodes the dynamically changing weight parameter. Third, an open theoretical task will be to generalize the model to explain other forms of Pavlovian-instrumental interactions, such as negative automaintenance and Pavlovian-instrumental transfer.

More broadly, our findings are consistent with the idea that agency is one factor that can mediate the trade-off between learning processes, which has important implications for understanding psychopathology. For example, many studies in both humans and animals have shown that controllability (or lack thereof) influences future instrumental responding. Learned helplessness, where the experience of uncontrollable punishments leads to diminished instrumental learning (for example, failure to learn to escape an electric shock; (Maier & Seligman, 1976), is hypothesized to be a model of, and has been linked to, symptoms of depression and anxiety (Mineka & Hendersen, 1985). The idea that inferences about controllability underlie learned helplessness has been incorporated into formal Bayesian models that share some properties with the model proposed in this paper (Lieder & Goodman, 2013).

In conclusion, we have shown how the framework of Bayesian model averaging can shed light on the cognitive mechanisms underlying Pavlovian misbehavior. Although the simple model studied in this paper is not a comprehensive theory of Pavlovian-instrumental interactions, it points towards one mechanism that is likely to play an important role in future, more comprehensive theories.

Funding
This work was supported by the National Institutes of Health (CRCNS 1R01MH109177), the Office of Naval Research (N00014-17-1-2984), and the Alfred P. Sloan Foundation.

Acknowledgements
The authors would like to thank Rebecca Hao for her help with the initial setup for this study.

References

Breland, K., & Breland, M. (1961). The misbehavior of organisms. *American Psychologist*, *16*(11), 681–684. http://doi.org/10.1037/h0040090

Cavanagh, J. F., Eisenberg, I., Guitart-Masip, M., Huys, Q., & Frank, M. J. (2013). Frontal Theta Overrides Pavlovian Learning Biases. *Journal of Neuroscience*, *33*(19), 8541–8548. http://doi.org/10.1523/JNEUROSCI.5754-12.2013

Courville, A. C., Daw, N. D., & Touretzky, D. S. (2006). Bayesian theories of conditioning in a changing world. *Trends in Cognitive Sciences*, *10*(7), 294–300. http://doi.org/10.1016/j.tics.2006.05.004

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, *8*(12), 1704–1711. http://doi.org/10.1038/nn1560

Dayan, P., Niv, Y., Seymour, B., & Daw, N. D. (2006). The misbehavior of value and the discipline of the will. *Neural Networks: the Official Journal of the International Neural Network Society*, *19*(8), 1153–1160. http://doi.org/10.1016/j.neunet.2006.03.002

Geman, S., Bienenstock, E., computation, R. D. N., 1992. (1992). Neural networks and the bias/variance dilemma. *MIT Press*, *4*(1), 1–58.

Germain, P., Bach, F., Lacoste, A., & Lacoste-Julien, S. (2016). PAC-Bayesian Theory Meets Bayesian Inference, 1884–1892.

Gershman, S. J. (2017). Deconstructing the human algorithms for exploration. *Cognition*, *173*, 34–42. http://doi.org/10.1016/j.cognition.2017.12.014

Gershman, S. J., Blei, D. M., & Niv, Y. (2010). Context, learning, and extinction. *Psychological Review*, *117*(1), 197–209. http://doi.org/10.1037/a0017808

Gigerenzer, G., & Brighton, H. (2009). Homo Heuristicus: Why Biased Minds Make Better Inferences. *Topics in Cognitive Science*, *1*(1), 107–143. http://doi.org/10.1111/j.1756-8765.2008.01006.x

Glaze, C. M., Filipowicz, A. L. S., Kable, J. W., Balasubramanian, V., & Gold, J. I. (2018). A bias–variance trade-off governs individual differences in on-line learning in an unpredictable environment. *Nature Human Behaviour*, 1–14. http://doi.org/10.1038/s41562-018-0297-4

Guitart-Masip, M., Chowdhury, R., Sharot, T., Dayan, P., Duzel, E., & Dolan, R. J. (2012a). Action controls dopaminergic enhancement of reward representations. *Proceedings of the National Academy of Sciences*, *109*(19), 7511–7516. http://doi.org/10.1073/pnas.1202229109

Guitart-Masip, M., Duzel, E., Dolan, R., & Dayan, P. (2014). Action versus valence in decision making. *Trends in Cognitive Sciences*, *18*(4), 194–202. http://doi.org/10.1016/j.tics.2014.01.003

Guitart-Masip, M., Huys, Q. J. M., Fuentemilla, L., Dayan, P., Duzel, E., & Dolan, R. J. (2012b). Go and no-go learning in reward and punishment: Interactions between affect and effect. *NeuroImage*, 1–13. http://doi.org/10.1016/j.neuroimage.2012.04.024

Hershberger, W. A. (1986). An approach through the looking-glass. *Animal Learning & Behavior*, *14*(4), 443–451. http://doi.org/10.3758/bf03200092

Huys, Q., & Dayan, P. (2009). A Bayesian formulation of behavioral control. *Cognition*, *113*(3), 314–328. http://doi.org/10.1016/j.cognition.2009.01.008

Lieder, F., Goodman, N. D, & Huys, Q. J. (2013). Learned helplessness and generalization. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 35. https://escholarship.org/uc/item/31362551.

MacKay, D. J. C. (2003). Information Theory, Inference and Learning Algorithms. Cambridge University Press.

Maier, S. F., & Seligman, M. E. (1976). Learned helplessness: Theory and evidence. *Journal of Experimental Psychology. General*, *105*(1), 3–46. http://doi.org/10.1037//0096-3445.105.1.3

Miller, S., & Konorski, J. (1928). On a particular form of conditioned reflex. *Journal of the Experimental Analysis of Behavior*, *12*(1), 187–189. http://doi.org/10.1901/jeab.1969.12-187.

Mineka, S., & Hendersen, R. W. (1985). Controllability and predictability in acquired motivation. *Annual Review of Psychology*, *36*(1), 495–529. http://doi.org/10.1146/annurev.ps.36.020185.002431

Moscarello, J. M., & Hartley, C. A. (2017). Agency and the Calibration of Motivated Behavior. *Trends in Cognitive Sciences*, *21*(10), 725–735. http://doi.org/10.1016/j.tics.2017.06.008

Mowrer, O.H., (1947). On the dual nature of learning—a re-interpretation of "conditioning" and "problem-solving." *Harvard Educational Review, 17,* 102-148.

Rescorla, R.A., & Solomon, R.L. (1967). Two-process learning theory: Relationships between Pavlovian conditioning and instrumental learning., *74*(3), 151–182.

Stephan, K. E., Penny, W. D., Daunizeau, J., Moran, R. J., & Friston, K. J. (2009). Bayesian model selection for group studies. *NeuroImage*, *46*(4), 1004–1017. http://doi.org/10.1016/j.neuroimage.2009.03.025

Tomov, M. S., Dorfman, H. M., & Gershman, S. J. (2018). Neural Computations Underlying Causal Structure Learning. *Journal of Neuroscience*, *38*(32), 7143–7157. http://doi.org/10.1523/JNEUROSCI.3336-17.2018

Wasserman, E. A., Franklin, S. R., & Hearst, E. (1974). Pavlovian appetitive contingencies and approach versus withdrawal to conditioned stimuli in pigeons. *Journal of Comparative and Physiological Psychology*, *86*(4), 616–627.

Williams, D. R., & Williams, H. (1969). Auto-maintenance in the pigeon: sustained pecking despite contingent non-reinforcement. *Journal of the Experimental Analysis of Behavior*, *12*(4), 511–520. http://doi.org/10.1901/jeab.1969.12-511