Computational Mechanisms Underlying the Influence of Agency on Learning

Computational Mechanisms Underlying the Influence of Agency on Learning

A dissertation presented

by

Hayley Dorfman

to

The Department of Psychology

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Psychology

Harvard University

Cambridge, Massachusetts

May 2019

Dissertation Advisor: Dr. Samuel Gershman Hayley Dorfman

Computational Mechanisms Underlying the Influence of Agency on Learning

Abstract

We live in an uncertain environment where making flexible predictions about the occurrence of positive and negative events is necessary for maximizing rewards, minimizing punishments, and guiding future behavior. Predictions are most accurate, and feedback most useful, when our own actions are responsible for the outcomes we receive. Both humans and animals exhibit a bias toward presuming control, or *agency*, over outcomes (Mineka, 1985), and extensive work suggests that beliefs about agency have substantial effects on how individuals learn from different types of outcomes. By utilizing computational models, neuroimaging, and flexible behavioral tasks, this dissertation investigated the behavioral and neurobiological pathways through which humans make inferences about hidden information and determined how these inferences influence learning processes.

We first found that inference about hidden agents modulated biased learning of positive and negative outcomes within the same individuals. A novel Bayesian model could account for this learning asymmetry, demonstrating a mechanistic framework for understanding how causal attributions contribute to learning (Paper 1). Next, we showed that RPE signals in the dorsal and ventral striatum were scaled by both subjective and model-derived beliefs about agency, but in opposite directions, providing preliminary evidence that striatal learning is gated by causal inference (Paper 2). Finally, we show that controllability arbitrates the use of a Pavlovian over an

instrumental learning system (Paper 3). Together, these results suggest that beliefs about agency are at least one factor that influences how we learn from feedback and how we decide the types of learning processes to utilize.

Table of Contents

Acknowledgments

First, I would like to thank my advisor, Dr. Sam Gershman, for his generosity and kindness, and for allowing me the opportunity to learn from him. I would also like to thank the members of my committee, Drs. Jill Hooley, Leah Somerville, and Mina Cikara, for not only their helpful comments and discussion on this project, but also for their support and guidance throughout my graduate career.

I want to also acknowledge members of the Cognitive Computational Neuroscience Lab for their helpful feedback, and my collaborators, Brent Hughes, Rahul Bhui, and Momchil Tomov, for taking the time to mentor me in new skills and being my sounding board for ideas and questions.

I want to extend a special thanks to Katie Insel, Alex Rodman, and Maheen Shermohammed for their unwavering support and genuine friendship. And to members of my cohort, whose curiosity and kindness are inspiring.

Finally, thanks to my family, for their encouraging pride, and to Ed, for being by my side the whole way.

This work was generously funded by the Sackler Scholar Programme in Psychobiology, the National Institutes of Health, the Office of Naval Research, the Alfred P. Sloan Foundation, and the Harvard Mind, Brain, Behavior Initiative.

Introduction

Background and Implications

We live in an uncertain environment where making flexible predictions about the occurrence of positive and negative events is necessary for maximizing rewards, minimizing punishments, and guiding future behavior. Predictions are most accurate, and feedback most useful, when our own actions are responsible for the outcomes we receive. Both humans and animals exhibit a bias toward presuming control, or *agency*, over outcomes (Mineka, Gunnar, & Champoux, 1985). Humans attribute agency to various inanimate objects or external forces (Gilbert & Wilson, 2009; Guthrie, 1995) and presume that others' behavior is goal-directed (Rosset, 2008). Not only is the presumption of agency over outcomes adaptive for behavior, but research also suggests that it is rewarding (Kool, Getz, & Botvinick, 2013; Leotti & Delgado, 2011).

The study of agency spans a wide range of different literatures: self-representation, action representation, psychopathology, social interactions, and cognitive biases, to name a few. As such, definitions of agency are varied. However, throughout this dissertation, we accept the definition provided by Schmidt and Heumüller of agency as a causal attribution process that can be limited to the self or extended to the other (Schmidt & Heumüller, 2010). We will use *sense of agency* to refer to the extent to which one attributes outcomes to a self-generated action. Some theories also distinguish between two subcomponents of the sense of agency: feelings of agency and judgments of agency (Synofzik, Vosgerau, & Newen, 2008). Feelings of agency refer to implicit, lower-level attributions (for example, judgments of motor-outcome contingencies), while judgments of agency refer to explicit, perceptible, "belief-like" instances of causal

attributions. All of the work presented here deals with agency judgments, and not lower-level predictions or inferences about agency that happen outside of conscious awareness.

One way in which beliefs about agency influence behavior is that they produce a suite of cognitive biases. Numerous studies have provided evidence for the prevalence of a self-serving bias, defined as the attribution of good outcomes to oneself and bad outcomes to external forces (Campbell & Sedikides, 1999; Hughes & Zaki, 2015). For example, people are more likely to think a third party influenced a gamble when the outcome was a loss instead of a win (Morewedge, 2009), more likely to take credit for positive as opposed to negative outcomes (Bradley, 1978), and demonstrate decreased intentional binding (the perception that an action is causal to an outcome if they occur closely together in time) for monetary losses compared to gains as well as negative versus positive affect cues (Takahata et al., 2012; Yoshie & Haggard, 2013). Similarly, optimistic and pessimistic biases can be manipulated by changes in outcome controllability. For example, greater perceived control is associated with increased optimism bias (Weinstein, 1980a), and this finding has also been shown in a large meta-analysis (Klein & Helweg-Larsen, 2002). Researchers have also looked at how agency influences a variety of outcomes, including sleep quality (Sanford, Yang, Wellman, Liu, & Tang, 2010) and pain perception (Akitsuki & Decety, 2009). Determinations of controllability have also been shown to have substantial effects on mental health and well-being. Theorists have argued that experiencing a lack of control in early environment can influence increased attribution to external causes, leading to a vulnerability for anxiety and depression (Chorpita & Barlow, 1998), and an extensive animal literature supports such claims (Mineka, 1982). For example, rhesus monkeys show signs of extreme stress and depression (Stroebel, 1969) and elevated cortisol levels (Hanson, 1976) when a lever that controlled the presentation of harmful stimuli was removed,

even when the stimuli were not presented again. In a seminal study by Alloy and Abramson, healthy participants exhibited an agency bias for desired outcomes, and a "non-agency" bias for undesired outcomes, while depressed participants showed no such bias (Alloy & Abramson, 1979). This work suggests that biased beliefs of control may be protective against depression, and that these cognitive distortions arise not solely from a belief that individuals have control over positive outcomes, but that negative outcomes can be attributed to someone or something outside of oneself.

Prior work on fear and aversive learning has attempted to pinpoint a mechanism for the development of psychopathology related to agency. Not only is there evidence that exposure to uncontrollable aversive events impairs subsequent avoidance learning in both animals (Maier & Seligman, 1976) and humans (Hartley, Gorun, Reddan, Ramirez, & Phelps, 2014), but exposure to controllable aversive events results in enhanced avoidance learning that persists even in future scenarios (Volpicelli, 1983). These results suggest that controllability over aversive outcomes, and not just instances of stress, determines to what extent individuals experience psychological consequences. While this traditional learning work has been imperative for progressing our understanding of adverse psychopathological outcomes, it has not been able to provide a cohesive framework for uniting the wide array of literatures on these topics. We propose the use of a modern reinforcement learning framework and computational modeling techniques in order to better formalize and probe how causal inference shapes learning.

The Neurobiology of Agency and Learning

Several brain regions have been associated with beliefs about agency, with many studies suggesting distinct neural correlates for self-agency compared to external-agency beliefs.

Numerous studies that have examined the neural correlates of beliefs about self-agency compared to external-agency have found evidence for the role of temporoparietal junction (TPJ), precuneus, and dorsomedial prefrontal cortex (dmPFC) in external-agency, and insular regions in self-agency (David, Newen, & Vogeley, 2008; Farrer & Frith, 2002; Sperduti, Delaveau, Fossati, & Nadel, 2011). In addition, a study using positron emission technology (PET) imaging has demonstrated lateralization of the inferior parietal lobule, finding activation of the left inferior parietal lobule when participants performed an action themselves, and activation of the right inferior parietal lobule when participants watched someone else move an object on a screen (Chaminade & Decety, 2002). Taken together, this work implicates two seemingly distinct groups of "self-referential" and "other-referential" regions in agency beliefs. Sensory motor regions, specifically the pre-supplementary motor areas (SMA), have also been implicated in tasks involving agency judgments. While some research has shown that the pre-SMA is related to both beliefs about self-agency (Miele, Wager, Mitchell, & Metcalfe, 2011) and external agency (Sperduti et al., 2011), lesion studies suggest that this region may be encoding intentional action more broadly. For example, patients with lesions to the pre-SMA often suffer from anarchic hand syndrome, where they describe uncontrollable hand movements as the hand having a "will of its own."

The primary cortical regions associated with causal inference do not seem to demonstrate much overlap with regions implicated in learning about outcomes. Previous work in animals has shown that reward prediction errors (RPEs), which represent the difference between expected and received reinforcement, are encoded by dopaminergic neurons (Schultz, Dayan, & Montague, 1997), and human neuroimaging studies find a consistent RPE signal in the ventral striatum (McClure, Berns, & Montague, 2003; O'Doherty, Dayan, Friston, Critchley, & Dolan,

2003). Human neuroimaging studies on the integration of reinforcement learning and subjective causal attributions are sparse. However, studies of goal-directed behavior in both humans and animals implicate medial prefrontal cortex and dorsal striatum (anterior caudate) (Balleine & Doherty, 2009), and rodent studies suggest that a vmPFC-striatal circuit is necessary for control over positive and negative feedback (Amat et al., 2014). A single study on self-serving bias found activation in the dorsal anterior cingulate and in the dorsal striatum (Seidel et al., 2010), suggesting a possible updating mechanism taking place during this type of bias.

Current Research

One challenge for studying these topics is that it is difficult to measure how people estimate the causes of outcomes. By utilizing computational models, neuroimaging, and flexible behavioral tasks, we can better measure and define the components of latent causal inference. This dissertation will explore behavioral and neurobiological pathways through which humans make inferences about hidden information and investigate how these inferences influence learning processes across three studies. In **Paper 1**, I hypothesized that outcome attributions to internal or external forces would influence to what extent individuals learn from negative and positive feedback, and that a novel computational model that could estimate causal beliefs would be able to account for learning behavior. **Paper 2** explored how beliefs about agency and the associated impact on learning asymmetries for positive and negative feedback are represented in the brain. By optimizing the task used in Study 1 for functional magnetic resonance imaging (fMRI), I aimed to replicate our previous behavioral findings, determine whether striatal prediction errors scale along with latent causal inference, and look at whether there are functional distinctions within sub-regions of the striatum for learning that is modulated by causal

inference. In order to further examine the precise learning processes that are influenced by outcome controllability and to integrate our previous work with the broader reinforcement learning literature, **Paper 3** seeks to provide a computational explanation for the arbitration between Pavlovian learning, which involves stimulus-outcome associations, and instrumental learning, which involves action-outcome associations. By manipulating the controllability of rewarding outcomes, I aimed to show that Pavlovian processes are favored when controllability is low, and instrumental processes are favored when controllability is high.

References

Amat, J., Christianson, J. P., Aleksejev, R. M., Kim, J., Richeson, K. R., Watkins, L. R., & Maier, S. F. (2014). Control over a stressor involves the posterior dorsal striatum and the act/outcome circuit. *European Journal of Neuroscience*, *40*(2), 2352–2358. http://doi.org/10.1111/ejn.12609

Balleine, B. W., & Doherty, J. P. O. A. (2009). Human and Rodent Homologies in Action Control: Corticostriatal Determinants of Goal-Directed and Habitual Action. *Neuropsychopharmacology*, *35*(1), 48–69. http://doi.org/10.1038/npp.2009.131

Bradley, G. W. (1978). Self-serving biases in the attribution process: A reexamination of the fact or fiction question. *Journal of Personality and Social Psychology*, *36*(1), 56–71. http://doi.org/10.1037/0022-3514.36.1.56

Campbell, W. K., & Sedikides, C. (1999). Self-threat magnifies the self-serving bias: A meta-analytic integration. *Review of General Psychology*, *3*(1), 23–43. http://doi.org/10.1037/1089-2680.3.1.23

Chaminade, T., & Decety, J. (2002). Leader or follower? Involvement of the inferior parietal lobule in agency. *Neuroreport*, *13*(15), 1975–1978.

Chorpita, B. F., & Barlow, D. H. (1998). The development of anxiety: the role of control in the early environment. *Psychological Bulletin*, *124*(1), 3–21.

David, N., Newen, A., & Vogeley, K. (2008). The "sense of agency" and its underlying cognitive and neural mechanisms. *Consciousness and Cognition*, *17*(2), 523–534. http://doi.org/10.1016/j.concog.2008.03.004

Farrer, C., & Frith, C. D. (2002). Experiencing Oneself vs Another Person as Being the Cause of an Action: The Neural Correlates of the Experience of Agency. *NeuroImage*, *15*(3), 596–603. http://doi.org/10.1006/nimg.2001.1009

Gilbert, D. T., & Wilson, T. D. (2009). Why the brain talks to itself: Sources of error in emotional prediction. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1521), 1335–1341. http://doi.org/10.1098/rstb.2008.0305

Guthrie, S. E. (1995). Faces in the Clouds. Oxford University Press.

Hartley, C. A., Gorun, A., Reddan, M. C., Ramirez, F., & Phelps, E. A. (2014). Stressor controllability modulates fear extinction in humans. *Neurobiology of Learning and Memory*, *113*(C), 149–156. http://doi.org/10.1016/j.nlm.2013.12.003

Hughes, B. L., & Zaki, J. (2015). The neuroscience of motivated cognition. *Trends in Cognitive Sciences*, *19*(2), 62–64. http://doi.org/10.1016/j.tics.2014.12.006

Kool, W., Getz, S. J., & Botvinick, M. M. (2013). Neural representation of reward probability: evidence from the illusion of control. *Journal of Cognitive Neuroscience*, *25*(6), 852–861. http://doi.org/10.1162/jocn_a_00369

Leotti, L. A., & Delgado, M. R. (2011). The Inherent Reward of Choice. *Psychological Science*, *22*(10), 1310–1318. http://doi.org/10.1177/0956797611417005

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

McClure, S. M., Berns, G. S., & Montague, P. R. (2003). Temporal prediction errors in a passive learning task activate human striatum. *Neuron*, *38*(2), 339–346.

Miele, D. B., Wager, T. D., Mitchell, J. P., & Metcalfe, J. (2011). Dissociating neural correlates of action monitoring and metacognition of agency. *Journal of Cognitive Neuroscience*, *23*(11), 3620–3636. http://doi.org/10.1162/jocn_a_00052

Mineka, Susan & Gunnar, Megan & Champoux, Maribeth. (1986). Control and Early Socioemotional Development: Infant Rhesus Monkeys Reared in Controllable versus Uncontrollable Environments. Child Development. 57. 1241. 10.2307/1130447.

Morewedge, C. K. (2009). Negativity bias in attribution of external agency. *Journal of Experimental Psychology. General*, *138*(4), 535–545. http://doi.org/10.1037/a0016796

O'Doherty, J. P., Dayan, P., Friston, K., Critchley, H., & Dolan, R. J. (2003). Temporal difference models and reward-related learning in the human brain. *Neuron*, *38*(2), 329–337.

Sanford, L. D., Yang, L., Wellman, L. L., Liu, X., & Tang, X. (2010). Differential effects of controllable and uncontrollable footshock stress on sleep in mice. *Sleep*, *33*(5), 621–630.

Schmidt, T., & Heumüller, V. C. (2010). Probability judgments of agency: Rational or irrational? *Consciousness and Cognition*, *19*(1), 1–11. http://doi.org/10.1016/j.concog.2010.01.004

Schultz, W., Dayan, P., & Montague, P. R. (1997). A Neural Substrate of Prediction and Reward. *Science (New York, NY)*.

Seidel, E.-M., Eickhoff, S. B., Kellermann, T., Schneider, F., Gur, R. C., Habel, U., & Derntl, B. (2010). Who is to blame? Neural correlates of causal attribution in social situations. *Social Neuroscience*, *5*(4), 335–350. http://doi.org/10.1080/17470911003615997

Sperduti, M., Delaveau, P., Fossati, P., & Nadel, J. (2011). Different brain structures related to self- and external-agency attribution: a brief review and meta-analysis. *Brain Structure & Function*, *216*(2), 151–157. http://doi.org/10.1007/s00429-010-0298-1

Synofzik, M., Vosgerau, G., & Newen, A. (2008). I move, therefore I am: A new theoretical framework to investigate agency and ownership. *Consciousness and Cognition*, *17*(2), 411–424. http://doi.org/10.1016/j.concog.2008.03.008

Takahata, K., Takahashi, H., Maeda, T., Umeda, S., Suhara, T., Mimura, M., & Kato, M. (2012). It's Not My Fault: Postdictive Modulation of Intentional Binding by Monetary Gains and Losses. *PLoS ONE*, *7*(12), e53421–8. http://doi.org/10.1371/journal.pone.0053421

Yoshie, M., & Haggard, P. (2013). Negative Emotional Outcomes Attenuate Sense of Agency over Voluntary Actions. *Current Biology*, *23*(20), 2028–2032. http://doi.org/10.1016/j.cub.2013.08.034

Paper 1: Causal Inference About Good and Bad Outcomes

Hayley M. Dorfman, Rahul Bhui, Brent L. Hughes, and Samuel J. Gershman

Published in: *Psychological Science* (2019), *In Press*

Abstract

People learn differently from good and bad outcomes. We argue that valence-dependent learning asymmetries are partly driven by beliefs about the causal structure of the environment. If hidden causes can intervene to generate bad (or good) outcomes, then a rational observer will assign blame (or credit) to these hidden causes, rather than to the stable outcome distribution. Thus, a rational observer should learn less from bad outcomes when these are likely to have been generated by a hidden cause, and the pattern should reverse when hidden causes are likely to generate good outcomes. To test this hypothesis, we conducted two experiments in which we explicitly manipulated the behavior of hidden agents ($N = 80$, $N = 255$). This gave rise to both kinds of learning asymmetries in the same paradigm, as predicted by a novel Bayesian model. These results provide a mechanistic framework for understanding how causal attributions contribute to biased learning.

People are motivated to maximize rewards and minimize punishments, but when updating their beliefs, they often weigh good and bad news differently. The nature of this differential weighting remains puzzling. In some cases, animals and humans attend more to bad events, and learn more rapidly from punishments compared to rewards (Baumeister, Bratslavsky, Finkenauer, & Vohs, 2001; Taylor, 1991). Similarly, some studies of reinforcement learning have found that learning rates are higher for negative than for positive prediction errors (Christakou et al., 2013; Gershman, 2015b; Niv, Edlund, Dayan, & O'Doherty, 2012). However,

other work has demonstrated the opposite pattern of results—greater learning for positive outcomes, not only in reinforcement learning tasks (Kuzmanovic, Jefferson, & Vogeley, 2016; Lefebvre, Lebreton, Meyniel, Bourgeois-Gironde, & Palminteri, 2017; Moutsiana, Charpentier, Garrett, Cohen, & Sharot, 2015), but also in procedural (Wachter, Lungu, Liu, Willingham, & Ashe, 2009) and declarative learning tasks (Eil & Rao, 2011; Sharot, Korn, & Dolan, 2011).

Here, we explore the hypothesis that the direction of valence-dependent learning asymmetries depends on *beliefs about the causal structure* of the environment. To provide some intuition, we borrow an example from Abramson, Seligman & Teasdale (1978): consider a researcher who receives a rejection for a manuscript submission. The researcher's inferences about the cause of that feedback will influence whether she modifies the paper or appeals the decision. If the researcher believes that she was rejected because the paper was bad, she will revise the paper and take this new information into consideration for future submissions. However, if she believes that the rejection was due to the opinion of an unfair reviewer, she will be less likely to update her beliefs about the quality of the paper. In other words, she will "explain away" the rejection as being due to a hidden cause (e.g., the reviewer's caustic temperament) rather than her own behavior.

Abramson, Seligman, and Teasdale argued that "failure means more than merely the occurrence of a bad outcome" (Abramson, Seligman, & Teasdale, 1978). Rather, attribution of negative outcomes to *oneself* is what constitutes failure. According to learned helplessness theory, individuals with an "optimistic explanatory style" tend to attribute negative events to external forces, whereas those with a "pessimistic explanatory style" believe that the causes of negative events are internal. On this view, optimistic and pessimistic cognitive biases might arise from both differing experiences of reinforcements and beliefs about their causes. Namely, both

the availability of rewards and punishments in the environment and the degree to which these consequences are attributed to oneself determine to what extent positive and negative outcomes influence learning.

Valence-dependent learning asymmetries are important because they may give rise to systematic biases with real-world consequences. On the one hand, learning more from positive outcomes can give rise to unrealistic optimism (Sharot et al., 2011) and risk-seeking behavior (Niv et al., 2012). On the other hand, learning more from negative outcomes can lead to unrealistic pessimism (Maier & Seligman, 1976) and risk aversion (Smoski et al., 2008). Thus, understanding the determinants of these asymmetries may provide insights into a wide range of behavioral phenomena and provide necessary information to curtail their harmful consequences.

One limitation of many past studies examining valence-dependent learning asymmetries is that they do not directly measure or control participants' beliefs about causal structure, and hence they are not ideal for testing our hypothesis. In the present paper, we report a more direct test by manipulating the causal structure of a reinforcement learning task to induce both positively-biased and negatively-biased learning asymmetries in the same participants. Participants were asked to choose between two options with unknown reward probabilities and were informed that an agent could silently intervene to change the outcome either positively (benevolent condition), negatively (adversarial condition), or randomly (neutral condition). Based on a Bayesian model of causal inference, we expected that participants would update their beliefs about the reward probabilities more from negative than positive outcomes in the benevolent condition. This is due to the fact that negative outcomes could not have been caused by an interfering external agent, but instead must have been a result of the participant's enacted choice (i.e. sampled from the option's reward distribution). Likewise, we expected that

participants would learn more from positive compared to negative outcomes in the adversarial condition. To examine the robustness and flexibility of our model, Experiment 2 explored a more realistic scenario where the probability of latent agent intervention was unknown to participants.

Experiment 1

In Experiment 1, we manipulated the causal structure underlying a reinforcement learning task, such that a hidden agent occasionally intervened to produce particular outcome types (good, bad, or random). This allowed us to test our primary hypothesis that positive outcomes would be weighed more heavily when the hidden agent was adversarial, whereas negative outcomes would be weighed more heavily when the hidden agent was benevolent. Importantly, if participants make no causal attributions in the task, then they should learn equally well from positive and negative outcomes in all of the three experimental conditions. However, any disproportionate learning of positive or negative outcomes can be attributed to the experimental manipulation of the causal structure. We formalized this hypothesis in terms of a Bayesian model that incorporates the underlying causal structure by rationally assigning credit to the different possible sources of feedback. This model describes participants' beliefs about latent agent interventions, while also providing a mechanistic account for how beliefs are formed and how they influence learning from positive and negative feedback.

Method

Participants

80 participants (25 female, 52 male, 3 unreported) from Amazon Mechanical Turk completed a two-alternative forced choice behavioral task. The sample size was chosen in order to exceed

sample sizes from previous, related work (Lefebvre et al., 2017; Sharot et al., 2011). Participants were excluded from analyses if they failed to choose the stimulus with the higher reward probability for > 60% of trials, leaving data from 72 participants (20 female, 49 male, 3 unreported) for subsequent analyses (90% of participants met the accuracy criterion). Participants gave informed consent, and the Harvard Committee on the Use of Human Subjects approved the experiment.

Procedure

Participants were instructed to imagine that they were mining for gold in the Wild West. On each trial, participants had to choose between one of two different colored mines by clicking on a button underneath the mine of their choice (Fig. 1.1b). After making a decision, participants received either a reward of gold or a loss of rocks (Fig. 1.1b). Each mine in a pair produced a reward with either 70% or 30% probability. Each reward yielded a small amount of real bonus money (5 cents) and each loss resulted in a subtraction of real bonus money (5 cents). Bonuses were summed, revealed, and paid out at the end of the task.

Participants completed three blocks of 50 trials (150 total trials) in different "mining territories" (Fig. 1.1a). Participants were instructed that different agents frequent each territory: a bandit will steal gold from the mines and replace it with rocks (adversarial condition), a tycoon will leave extra gold in the mines (benevolent condition), and a neutral sheriff will try to redistribute gold and rocks in the mines (neutral condition). Participants completed each of the three conditions once, in randomized order. The agents intervened on 30% of the trials, and participants were told this proportion explicitly at the start of the task, though they did not know unambiguously whether the agent intervened on any particular trial. While the underlying reward

distributions (i.e., absent intervention) for the mines were 70% or 30%, the hidden agent intervened on 30% of trials (or 15 out of 50 trials). For example, the benevolent intervention produces rewards on 15 out of 50 trials, the adversarial intervention produces losses on 15 out of 50 trials, and the random intervention produces either losses or rewards for 15 out of 50 trials. After feedback on each trial, participants were asked whether they believed the outcome they received was a result of hidden agent intervention (binary response of "Yes" or "No"; Fig. 1.1b).

To ensure that participants understood task instructions, they were asked comprehension questions that required correct answers before proceeding.

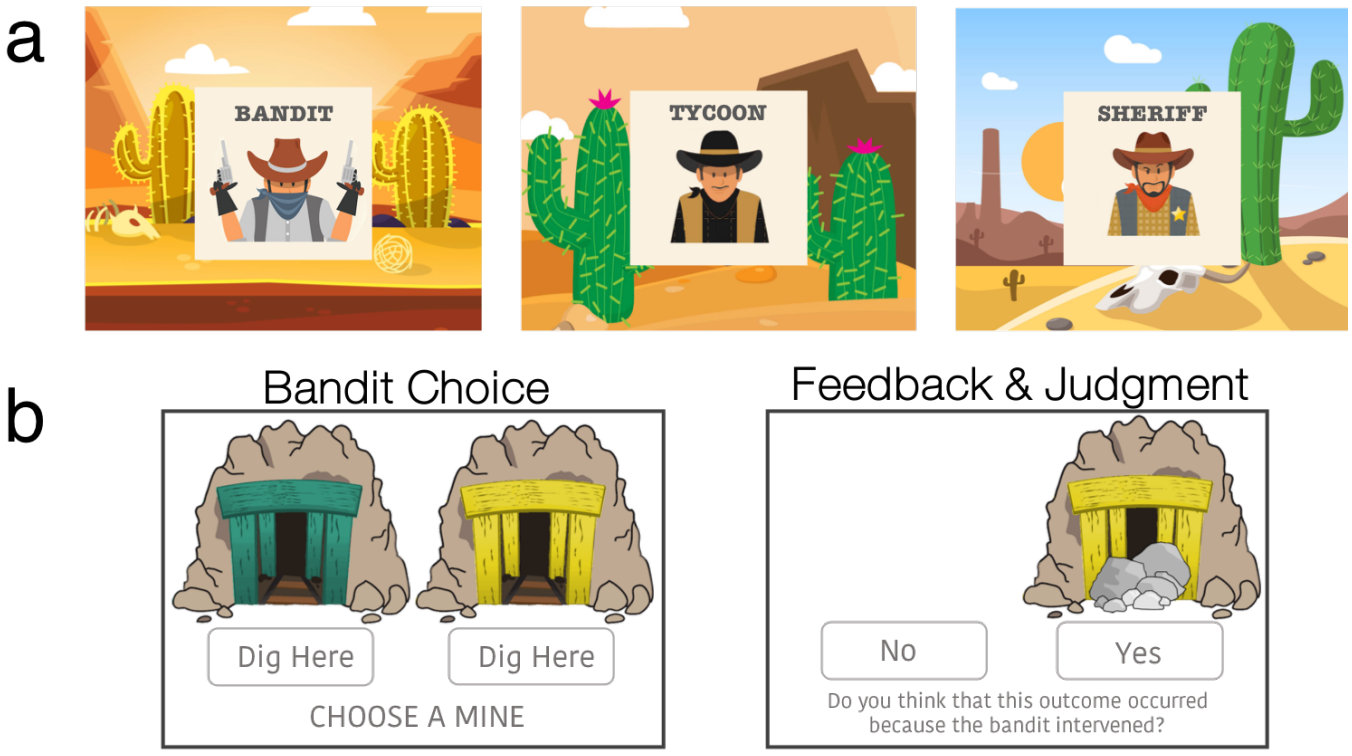


Figure 1.1. Behavioral task schematic. (a) Participants are told the latent agent intervention condition (adversarial, benevolent, or neutral) at the start of each block. (b) Participants make a choice on a two-armed bandit. They then receive feedback and judge whether or not they believe that the latent agent intervened on the current trial.

Bayesian reinforcement learning model

The problem facing the participant during the task is to choose the option yielding the highest reward. Because she does not know the reward probabilities of the two arms, she must estimate them from experience, while taking into account possible intervention from the hidden agent. We developed a Bayesian reinforcement learning model that jointly infers the hidden agent interventions and the reward probabilities. Here, we summarize the model (see Supplemental Materials for a full mathematical description.).

After choosing an action and observing reward r_t on trial t, the participant updates her estimate of the action's intrinsic reward probability θ_t according to a reinforcement learning equation that depends on inferences about latent causes: $\theta_{t+1} = \theta_t + \alpha_t(r_t - \theta_t)$, where α_t is a learning rate. The learning rate changes across trials depending on the posterior probability of the hidden agent intervention, as computed by Bayes' rule. When the posterior probability is high, the learning rate is low. Intuitively, the model predicts that the participant should suspend learning about the reward probabilities when she believes that the outcome was generated by an external force. Although the model is derived from Bayesian principles, at a mechanistic level it closely resembles standard reinforcement learning models that update reward predictions based on prediction errors. Like other Bayesian reinforcement learning models, the dynamic learning rate is derived from probabilistic assumptions about the environment (e.g., (Frank, Doll, Oas-Terpstra, & Moreno, 2009; Gershman, 2015a; Gershman & Niv, 2015) rather than left as a free parameter. However, we enrich typical reinforcement learning models by rationally assigning credit to different possible sources of feedback. In other words, the learning rate in the Bayesian model is calculated by integrating one's cumulative, past beliefs about intervention into one's value estimate of a particular choice.

Critically, the learning rate exhibits asymmetries depending on whether the hidden agent tends to produce positive or negative outcomes. For example, when the agent is adversarial, positive outcomes can only be generated from the intrinsic reward probabilities, whereas negative outcomes can be generated by either the hidden agent or the intrinsic reward probabilities. Consequently, negative outcomes are less informative about the reward probabilities in this scenario, inducing a lower learning rate.

The Bayesian model was fit using maximum a posteriori estimation with empirical priors based on previous research (Gershman, 2016). We compute a posterior over the underlying parameters, and then input the expected values into a softmax function to model choice probabilities, with a response stochasticity (inverse temperature) parameter and a "stickiness" parameter to capture choice autocorrelation (Gershman, Pesaran, & Daw, 2009).

A valence-dependent learning asymmetry is built into the structure of the Bayesian model. Thus, the model itself cannot be used to test for the existence of such an asymmetry. To provide evidence for asymmetric learning, we also fit a reinforcement learning (RL) model in which we modeled separate, fixed learning rates for positive and negative outcomes in each of the three experimental conditions (six learning rates total). This RL model is “heuristic” in the sense that it characterizes, but does not explain mechanistically, learning rate asymmetries in our task. Importantly, this model allows for differential weighting of positively and negatively valenced outcomes without taking into account hidden agent interventions.

We used random-effects Bayesian model selection (Rigoux, Stephan, Friston, & Daunizeau, 2014; Stephan, Penny, Daunizeau, Moran, & Friston, 2009) to compare models. This procedure treats each participant as a random draw from a population-level distribution over models, which it estimates from the sample of model evidence values for each model. We used

the Laplace approximation of the log marginal likelihood to obtain the model evidence values. For our model comparison metric, we report 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.

Results

Behavioral analyses

As a preliminary manipulation check, we verified that participants' beliefs about hidden causes varied with the outcome valence in a condition-specific manner (Fig. 1.2a). Participants were more likely to believe that a hidden cause resulted in negative, as opposed to positive outcomes overall, $t(71) = 16.82$, $p < 0.0001$, $d = -0.32$, 95% CI [-0.55 -0.08]. Importantly, participants were more likely to believe that the hidden agent had intervened after negative compared to positive outcomes in the adversarial condition, $t(71) = 66.24$, $p < 0.0001$, $d = -0.28$, 95% CI [-0.51 -0.04], and after positive compared to negative outcomes in the benevolent condition, $t(71) = -71.35$, $p < 0.0001$, $d = -0.99$, 95% CI [-1.23 -0.74]. Participants were also slightly more likely to be believe that the hidden agent had intervened after negative outcomes in the neutral condition, $t(71) = 6.83$, $p < 0.0001$, $d = -0.31$, 95% CI [-0.55 -0.08]. We will revisit this effect in the context of our computational model.

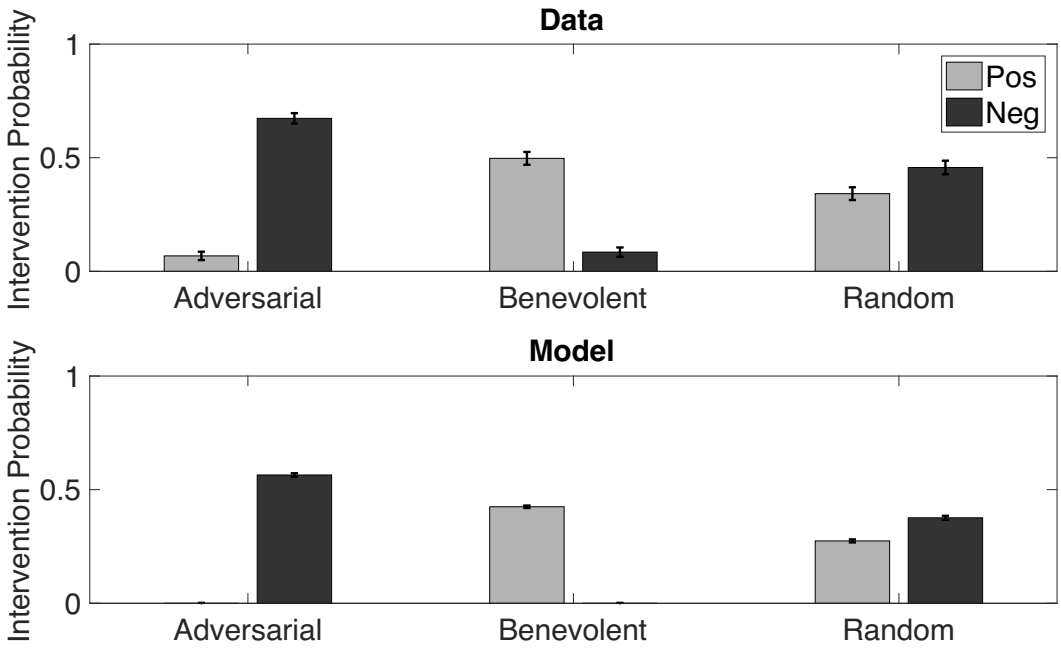


Figure 1.2. Beliefs about hidden agent intervention in Experiment 1. (a) Average belief about hidden agent intervention by condition and feedback type. Intervention probability was calculated by taking the mean of each participant's guess of whether or not the latent agent caused a given outcome for each trial. (b) Bayesian model predictions. Across-subject standard error of the mean (SEM) is shown for both panels.

Computational modeling

To characterize the effects of outcome valence and agent type on learning, we first fit a RL model with six separate learning rates. As shown in Fig. 1.3a, participants generally learned more from positive than from negative outcomes across all conditions, $t(71) = 5.56$, $p < 0.0001$, $d = 0.66$, 95% CI [0.32 0.99]. By treating the positivity bias in the neutral condition, $t(71) = 3.08$, $p < 0.003$, $d = 0.36$, 95% CI [0.03 0.70], as a participant-specific baseline and subtracting it from the other conditions, we obtained a *relative* measure of learning rates for the adversarial and benevolent conditions (Fig. 1.3b), revealing an underlying sensitivity to condition and valence. A 2 x 2 (adversarial vs. benevolent and positive vs. negative) repeated measures ANOVA on relative learning rates revealed no significant main effects ($p = 0.57$ for condition, $p = 0.84$ for valence) but a significant interaction [$F(1,71) = 4.91$, $p < 0.05$]. Consistent with our hypothesis,

the learning rate advantage for positive vs. negative outcomes reverses depending on the causal structure of the task.

Although the interaction effect is significant, the effects are small and noisy because the model is over-parametrized: by modeling learning rates separately for each condition and outcome valence, it cannot capture a common learning mechanism and thus has a small amount of data from which to estimate each parameter. We therefore developed a Bayesian reinforcement learning model that makes explicit the common learning mechanism. Learning rates in the Bayesian model are determined entirely by the causal structure, which is known to the participant. The only free parameters in the model are those governing the choice policy (response stochasticity and stickiness). We fixed the prior probability of hidden agent intervention in the Bayesian model at 30% to replicate the instructions that participants received in the task (a variant of the model in which this probability was treated as a free parameter for each participant yielded a value very close to the ground truth: mean = 0.296, *SEM* = 0.038).

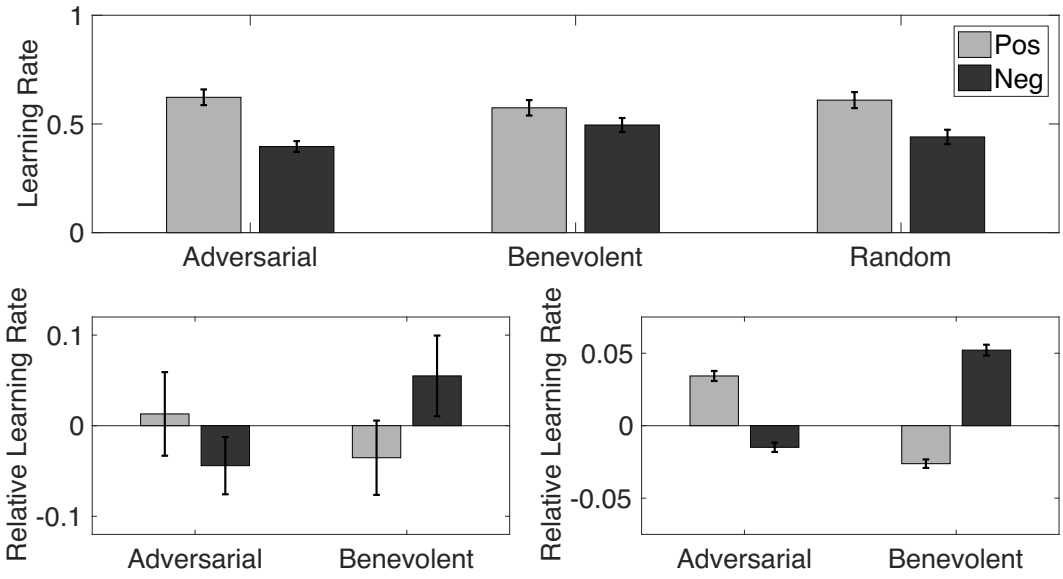


Figure 1.3. Learning rates. (a) Reinforcement learning (RL) model with separate learning rates for each condition (adversarial, benevolent, neutral) and outcome valence (positive, negative). Across all conditions, there is a tendency to learn more from positive than from negative outcomes. (b) Subtracting the learning rates for the neutral condition (which serves as a participant-specific baseline) from the other conditions, a cross-over interaction between condition and outcome valence is revealed. (c) Relative learning rates from the Bayesian model, averaged across trials.

As expected, the Bayesian model shows a strong interaction between condition and outcome valence (Fig. 1.3c), a direct consequence of causal inference. To bolster our claim that causal inference predicts the valence-dependent learning rate asymmetry, we examined the relationship between intervention judgments and learning rates (derived from the Bayesian model), finding that learning rates were significantly lower for trials on which participants believed that the hidden agent intervened, compared to trials on which they believed that the hidden agent did not intervene, $t(71) = -6.94$, $p < 0.0001$, $d = -0.82$; Fig. 1.4a.

Quantitative metrics also supported the Bayesian model. First, the model can predict intervention judgments even though it was not fit to these judgments: a signed rank test between the participants' guesses about intervention and the model's posterior over intervention showed a significant median point-biserial correlation ($r_{pb} = 0.45$, $p < 0.0001$). Second, the Bayesian model

received unequivocally stronger support than the reinforcement learning model with six learning rates according to a random-effects model selection procedure ($PXP = 0.97$ and $PXP < 0.001$). This model was also strongly favored over the Bayesian model with a free intervention probability parameter ($PXP < 0.001$) and a non-Bayesian reinforcement learning model with separate learning rates for positive and negative feedback ($PXP = 0.03$).

Discussion

Participants in Experiment 1 demonstrated asymmetric learning from positive and negative outcomes that reversed depending on the nature of a hidden intervening agent: when the hidden agent intervened to produce negative outcomes, learning was greater for positive outcomes, and when the hidden agent intervened to produce positive outcomes, learning was greater for negative outcomes. A Bayesian model captured this pattern and could also accurately predict participant's trial-by-trial judgments about interventions. As predicted by the model, learning rates were lower when participants believed that the hidden agent intervened. These results support our hypothesis that causal inference plays a central role in determining valence-dependent learning asymmetries.

Note that our data are not consistent with a model in which subjects follow a simple rule of ignoring negative feedback in the adversarial environment and positive feedback in the benevolent environment. If they were in fact following such a rule, then we would expect learning rates in those cases to be 0, whereas in fact they are significantly greater than 0. The Bayesian model captures the differential sensitivity to positive and negative feedback in a more graded manner than a simple rule-based model.

Experiment 2

One of the broader questions motivating this research is how the environment shapes learning rate asymmetries. We addressed this question in Experiment 2 by creating a subtle ambiguity in our experimental task: instead of informing participants of the exact intervention probability, we simply told them that hidden agents occasionally intervene. We reasoned that this ambiguity more directly reflects experiences in the real world in which probabilities for interventions and outcomes are often unknown and would be resolved by participants in different ways depending on their prior expectations.

Method

Participants

299 participants ("Sample A": N = 110 [49 female, 56 male, 5 unreported]; "Sample B": N = 194 [90 female, 96 male, 8 unreported]) recruited from Amazon Mechanical Turk completed a two-alternative forced choice behavioral task. Sample B was collected as part of a pre-registered replication, though for the purposes of these analyses we have aggregated the two samples (see Supplemental Materials for further information on the pre-registered replication. Registration details can be found on the Open Science Framework: https://osf.io/3htpj/). Participants were excluded from model fitting if they did not choose the stimulus with the higher reward probability for > 60% of trials. 85.3% of all participants met the accuracy criterion (86.4% [Sample A], 84.7% [Sample B]). Participants were also excluded if they did not properly respond to an attention check question (N = 6). We included data from 255 participants in the model fits (N = 95 for Sample A; N = 160 for Sample B). Participants gave informed consent, and the Harvard Committee on the Use of Human Subjects approved the experiment.

Procedure

Behavioral task procedures were identical to those outlined in Experiment 1, except that participants were told that the hidden agents would intervene "sometimes." Actual intervention remained fixed at 30% (15 of 50 trials per block).

Computational Model

Because participants were not told the intervention probability, we explored models that either estimated the probability directly (the "adaptive Bayesian" model) or treated it as a free parameter (the "fixed Bayesian" model). In addition, we fit a model in which the intervention probability was derived empirically by simply averaging the binary intervention judgments. We refer to this model as the "empirical Bayesian" model.

Results

Behavioral analyses

Replicating the results of Experiment 1, participants believed that the hidden agent caused negative outcomes more often than positive outcomes across all conditions, $t(254) = 6.26$, $p < 0.0001$, $d = -0.06$, 95% CI [-0.18, 0.06], and there was a significant difference between belief in the hidden agent for good outcomes in the benevolent and neutral conditions, and for bad outcomes in the adversarial and neutral conditions, $t(254) = 7.03$, $p < 0.0001$, $d = 0.29$, 95% CI

[0.17, 0.41].

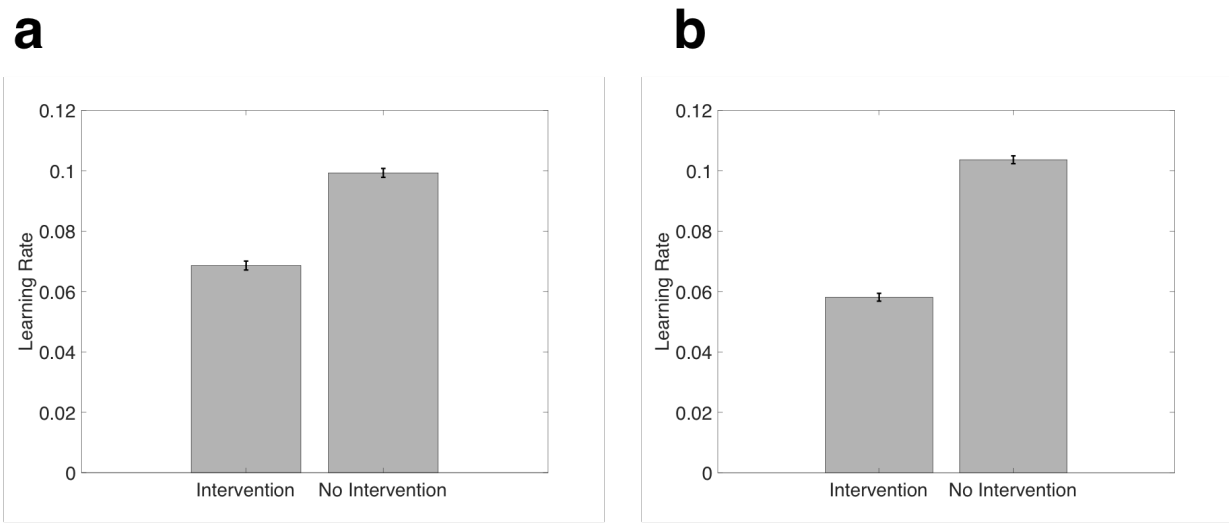


Figure 1.4. Learning rates for trials where participants did or did not believe the outcome was a result of hidden agent intervention for (a) Experiment 1 and (b) Experiment 2. Learning rates were derived from the Bayesian model (Experiment 1) and the empirical Bayesian model (Experiment 2).

Computational Modeling

Model comparison overwhelmingly (PXP > 0.999) supported the "empirical Bayesian" model (in which the intervention probability is derived from the binary intervention judgments) compared to a more sophisticated "adaptive Bayesian" model (which estimated the intervention probability from experience) and a "fixed Bayesian" model which treated the intervention probability as a free parameter.

Once again, we found that participants had significantly higher learning rates for positive outcomes compared to negative outcomes, $t(254) = 4.73$, $p < 0.0001$, $d = -0.82$, 95% CI [-0.95, -0.69]. Further replicating our results from Experiment 1, we also found that learning rates were significantly lower for trials where participants believed that the hidden agent intervened, compared to trials where they believed that the hidden agent did not intervene, $t(252) = 16.77$, $p < 0.0001$, $d = -0.80$, 95% CI [-0.93, -0.67]; Fig. 1.4b].

A signed rank test between the participants' actual guess about intervention and the intervention predicted by the model showed a significant median point-biserial correlation (r_{pb} = 0.55, $p < 0.0001$), demonstrating that intervention judgments can be accurately predicted by the adaptive Bayesian model.

Discussion

By modifying the behavioral task and the computational model to include an unknown probability of hidden agent intervention, we were able to gain insight into individual differences in prior expectations that govern valence-dependent learning asymmetries. First, a version of the Bayesian model that derived the intervention probability from the average of participants' binary judgments was favored by model selection among the models we considered. We conjecture that our task taps into prior expectations about the nature and frequency of hidden agents, possibly formed over a lifetime of learning.

General Discussion

Across two experiments, we found that the direction of valence-dependent learning asymmetries could be influenced by manipulating beliefs about causal structure. Specifically, participants learned more from positive than from negative outcomes when hidden agents intervened adversarially, and conversely, learned more from negative than from positive outcomes when hidden agents intervened benevolently. A Bayesian model explained the complete pattern of asymmetries, and an extension of the model that inferred the probability of hidden agent intervention could capture performance in the more complex scenario in which the intervention probability was unknown.

Our findings are consistent with the long-standing idea that optimistic biases are not exclusively a consequence of increased salience of positive outcomes, but also involve external attribution of negative outcomes (Miller & Ross, 1975). With replication across two independent samples, people displayed a generalized tendency to attribute positive outcomes to themselves and negative outcomes to others. Numerous studies have provided evidence for the prevalence of a self-serving bias (the attribution of good outcomes to oneself and bad outcomes to external forces) (Campbell & Sedikides, 1999; Hughes & Zaki, 2015). For example, people are more likely to think a third party influenced a gamble when the outcome was a loss instead of a win (Morewedge, 2009), more likely to take credit for positive as opposed to negative outcomes (Bradley, 1978), and demonstrate decreased intentional binding for monetary losses compared to gains as well as negative versus positive affect cues (Takahata et al., 2012; Yoshie & Haggard, 2013). Our computational model does not take into account a self-preservation bias, but removing individual variance by subtracting the bias in the neutral condition from the other two conditions results in our hypothesized asymmetry. In addition, our data show that participants have higher learning rates for positive outcomes. This discrepancy between learning rates for good and bad news is consistent with the well-studied phenomenon of an inherent optimism bias (Weinstein, 1980a). Therefore, it may be more difficult for individuals to discount their agency over rewards, even when they are provided with explicit instructions about the structure of the environment.

While our model can account for our experimental findings and the related phenomena reviewed above, we have not demonstrated that it provides a comprehensive account of optimism bias in general. It is difficult to attribute optimism bias to causal structure without knowing (or manipulating) subjects' structural beliefs, which was the starting point of the present paper.

Nonetheless, it is possible to offer some speculations (Gershman, 2018). If a person has a strong belief in their self-efficacy, then observing failure will favor the hypothesis that a latent cause was responsible, resulting in less updating compared to if success was observed. One suggestive source of data comes from studies of psychiatric disorders, where beliefs about self-efficacy and agency are disrupted.

Previous work has shown that learning asymmetries are associated with depression, anhedonia, and pessimism/optimism. For example, research demonstrates that patients with depression and anhedonia exhibit blunted learning for both rewards and punishments (Chase et al., 2010), and depressed participants accurately recall negative outcomes while healthy participants underestimate the frequency of negative outcomes (Nelson & Craighead, 1977). It should be noted, however, that other studies have also found no relationship between optimism and asymmetric updating (Stankevicius, Huys, Kalra, & Seriès, 2014). Our findings suggest that optimistic/pessimistic traits should depend on the interaction between imbalanced learning and beliefs about agency. We propose that a latent factor, such as agency inference, may be mediating inconsistent findings in the literature regarding imbalanced learning for positive and negative outcomes. This idea is supported by research that shows that optimistic and pessimistic biases can be manipulated by changes in outcome controllability. For example, greater perceived control is associated with increased optimism bias (Weinstein, 1980b), and this finding has also been shown in a large meta-analysis (Klein & Helweg-Larsen, 2002). In a seminal study by Alloy and Abramson, healthy participants exhibited an agency bias for desired outcomes, and a "non-agency" bias for undesired outcomes, while depressed participants showed no such bias (Alloy & Abramson, 1979). This work suggests that biased beliefs of control may be protective against depression, and that these cognitive distortions arise not solely from a belief that

individuals have control over positive outcomes, but that negative outcomes can be attributed to someone or something outside of oneself (though see(Msetfi, Murphy, Simpson, & Kornbrot, 2005), for evidence that abnormal beliefs about control in depression may be due to an impairment in contextual processing).

Conclusion

In sum, we provide evidence that valence-dependent learning asymmetries arise from causal inference over hidden agents. This idea, formalized in a simple Bayesian model, was able to quantitatively and qualitatively account for both choices and intervention judgments. An important task for future work will be to understand the limits of this framework: to what extent can we understand self-serving biases, learned helplessness, and other related behavioral phenomena in terms of a common computational mechanism? More generally, the real world is typically less well-behaved than the idealized experimental scenarios studied in this paper; people constantly face causally complex and ambiguous inferential problems, where simple attributions to "good" and "bad" latent agents may not be applicable. We foresee an exciting challenge in extending the Bayesian framework to tackle these more realistic settings.

Acknowledgments

H.M. Dorfman was supported by the Sackler Scholar Programme in Psychobiology. S.J. Gershman was supported by the National Institutes of Health (CRCNS R01-1207833) and Office of Naval Research (N00014-17-1-2984). R. Bhui was supported by the Harvard Mind Brain Behavior Initiative.

References

Abramson, L. Y., Seligman, M. E., & Teasdale, J. D. (1978). Learned helplessness in humans: Critique and reformulation. *Journal of Abnormal Psychology*, *87*(1), 49–74. http://doi.org/10.1037//0021-843x.87.1.49

Alloy, L. B., & Abramson, L. Y. (1979). Judgment of contingency in depressed and nondepressed students: Sadder but wiser? *Journal of Experimental Psychology. General*, *108*(4), 441–485. http://doi.org/10.1037/0096-3445.108.4.441

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology*, *5*(4), 323–370. http://doi.org/10.1037//1089-2680.5.4.323

Bradley, G. W. (1978). Self-serving biases in the attribution process: A reexamination of the fact or fiction question. *Journal of Personality and Social Psychology*, *36*(1), 56–71. http://doi.org/10.1037/0022-3514.36.1.56

Campbell, W. K., & Sedikides, C. (1999). Self-threat magnifies the self-serving bias: A meta-analytic integration. *Review of General Psychology*, *3*(1), 23–43. http://doi.org/10.1037/1089-2680.3.1.23

Chase, H. W., Frank, M. J., Michael, A., Bullmore, E. T., Sahakian, B. J., & Robbins, T. W. (2010). Approach and avoidance learning in patients with major depression and healthy controls: relation to anhedonia. *Psychological Medicine*, *40*(3), 433–440. http://doi.org/10.1017/S0033291709990468

Christakou, A., Gershman, S. J., Niv, Y., Simmons, A., Brammer, M., & Rubia, K. (2013). Neural and Psychological Maturation of Decision-making in Adolescence and Young Adulthood. *Journal of Cognitive Neuroscience*, *25*(11), 1807–1823. http://doi.org/10.1162/jocn_a_00447

Eil, D., & Rao, J. M. (2011). The Good News-Bad News Effect: Asymmetric Processing of Objective Information about Yourself. *American Economic Journal: Microeconomics*, *3*(2), 114–138. http://doi.org/10.1257/mic.3.2.114

Frank, M. J., Doll, B. B., Oas-Terpstra, J., & Moreno, F. (2009). Prefrontal and striatal dopaminergic genes predict individual differences in exploration and exploitation. *Nature Neuroscience*, *12*(8), 1062–1068. http://doi.org/10.1038/nn.2342

Gershman, S. J. (2015a). A Unifying Probabilistic View of Associative Learning. *PLoS Computational Biology*, *11*(11), e1004567–. http://doi.org/10.1371/journal.pcbi.1004567

Gershman, S. J. (2015b). Do learning rates adapt to the distribution of rewards? *Psychonomic Bulletin & Review*, *22*(5), 1320–1327. http://doi.org/10.3758/s13423-014-0790-3

Gershman, S. J. (2016). Empirical priors for reinforcement learning models. *Journal of Mathematical Psychology*, *71*, 1–6. http://doi.org/10.1016/j.jmp.2016.01.006

Gershman, S. J. (2018). How to never be wrong, 1–16. http://doi.org/10.3758/s13423-018-1488-8

Gershman, S. J., & Niv, Y. (2015). Novelty and Inductive Generalization in Human Reinforcement Learning. *Topics in Cognitive Science*, *7*(3), 391–415. http://doi.org/10.1111/tops.12138

Gershman, S. J., Pesaran, B., & Daw, N. D. (2009). Human reinforcement learning subdivides structured action spaces by learning effector-specific values. *The Journal of Neuroscience : the Official Journal of the Society for Neuroscience*, *29*(43), 13524–13531. http://doi.org/10.1523/JNEUROSCI.2469-09.2009

Hughes, B. L., & Zaki, J. (2015). The neuroscience of motivated cognition. *Trends in Cognitive Sciences*, *19*(2), 62–64. http://doi.org/10.1016/j.tics.2014.12.006

Klein, C. T. F., & Helweg-Larsen, M. (2002). Perceived Control and the Optimistic Bias: A Meta-Analytic Review. *Psychology & Health*, *17*(4), 437–446. http://doi.org/10.1080/0887044022000004920

Kuzmanovic, B., Jefferson, A., & Vogeley, K. (2016). The role of the neural reward circuitry in self-referential optimistic belief updates. *NeuroImage*, *133*, 151–162.

Lefebvre, G., Lebreton, M., Meyniel, F., Bourgeois-Gironde, S., & Palminteri, S. (2017). Behavioural and neural characterization of optimistic reinforcement learning. *Nature Human Behaviour*, *1*, 0067. http://doi.org/10.1038/s41562-017-0067

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, D. T., & Ross, M. (1975). Self-serving biases in the attribution of causality: Fact or fiction? *Psychological Bulletin*, *82*(2), 213–225. http://doi.org/10.1037/h0076486

Morewedge, C. K. (2009). Negativity bias in attribution of external agency. *Journal of Experimental Psychology. General*, *138*(4), 535–545. http://doi.org/10.1037/a0016796

Moutsiana, C., Charpentier, C. J., Garrett, N., Cohen, M. X., & Sharot, T. (2015). Human Frontal-Subcortical Circuit and Asymmetric Belief Updating. *The Journal of Neuroscience : the Official Journal of the Society for Neuroscience*, *35*(42), 14077–14085. http://doi.org/10.1523/JNEUROSCI.1120-15.2015

Msetfi, R. M., Murphy, R. A., Simpson, J., & Kornbrot, D. E. (2005). Depressive Realism and Outcome Density Bias in Contingency Judgments: The Effect of the Context and Intertrial Interval. *Journal of Experimental Psychology. General*, *134*(1), 10–22. http://doi.org/10.1037/0096-3445.134.1.10

Nelson, R. E., & Craighead, W. E. (1977). Selective recall of positive and negative feedback, self-control behaviors, and depression. *Journal of Abnormal Psychology*, *86*(4), 379–388.

Niv, Y., Edlund, J. A., Dayan, P., & O'Doherty, J. P. (2012). Neural Prediction Errors Reveal a Risk-Sensitive Reinforcement-Learning Process in the Human Brain. *Journal of Neuroscience*, *32*(2), 551–562. http://doi.org/10.1523/JNEUROSCI.5498-10.2012

Rigoux, L., Stephan, K. E., Friston, K. J., & Daunizeau, J. (2014). Bayesian model selection for group studies - Revisited. *NeuroImage*, *84*(C), 971–985. http://doi.org/10.1016/j.neuroimage.2013.08.065

Sharot, T., Korn, C. W., & Dolan, R. J. (2011). How unrealistic optimism is maintained in the face of reality. *Nature Neuroscience*, *14*(11), 1475–1479. http://doi.org/10.1038/nn.2949

Smoski, M. J., Lynch, T. R., Rosenthal, M. Z., Cheavens, J. S., Chapman, A. L., & Krishnan, R. R. (2008). Decision-making and risk aversion among depressive adults. *Journal of Behavior Therapy and Experimental Psychiatry*, *39*(4), 567–576. http://doi.org/10.1016/j.jbtep.2008.01.004

Stankevicius, A., Huys, Q. J. M., Kalra, A., & Seriès, P. (2014). Optimism as a Prior Belief about the Probability of Future Reward. *PLoS Computational Biology*, *10*(5), e1003605. http://doi.org/10.1371/journal.pcbi.1003605

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

Takahata, K., Takahashi, H., Maeda, T., Umeda, S., Suhara, T., Mimura, M., & Kato, M. (2012). It's Not My Fault: Postdictive Modulation of Intentional Binding by Monetary Gains and Losses. *PLoS ONE*, *7*(12), e53421–8. http://doi.org/10.1371/journal.pone.0053421

Taylor, S. E. (1991). Asymmetrical effects of positive and negative events: the mobilization-minimization hypothesis. *Psychological Bulletin*, *110*(1), 67–85.

Wachter, T., Lungu, O. V., Liu, T., Willingham, D. T., & Ashe, J. (2009). Differential Effect of Reward and Punishment on Procedural Learning. *Journal of Neuroscience*, *29*(2), 436–443. http://doi.org/10.1523/JNEUROSCI.4132-08.2009

Weinstein, N. D. (1980a). Unrealistic optimism about future life events. *Journal of Personality and Social Psychology*, *39*(5), 806–820. http://doi.org/10.1037/0022-3514.39.5.806

Weinstein, N. D. (1980b). Unrealistic optimism about future life events. *Journal of Personality and Social Psychology*, *39*(5), 806–820. http://doi.org/10.1037//0022-3514.39.5.806

Yoshie, M., & Haggard, P. (2013). Negative Emotional Outcomes Attenuate Sense of Agency over Voluntary Actions. *Current Biology*, *23*(20), 2028–2032. http://doi.org/10.1016/j.cub.2013.08.034

Paper 2: Causal Inference Gates Learning in the Striatum

Hayley M. Dorfman, Momchil Tomov, Bernice Chung, Dennis Clarke, Brent L. Hughes*, and Samuel J. Gershman*

*Equal Contribution

(In Preparation)

Abstract

Learning from positive and negative outcomes can be biased by beliefs about outcome controllability, but little is known about the neural systems underlying the interaction between controllability and valence-dependent learning asymmetries. In this study, participants ($N = 36$) were scanned using functional magnetic resonance imaging (fMRI) while completing a behavioral task that manipulated beliefs about causal structure underlying a two-armed bandit task. Replicating previous work, the direction of valence-dependent learning asymmetry was dependent on beliefs about causal structure, as predicted by a Bayesian reinforcement learning model. Specifically, participants learned more from positive reward prediction errors (RPEs) than from negative RPEs when they believed that an adversarial hidden agent may have intervened, whereas they learned more from negative RPEs when they believed that a benevolent hidden agent may have intervened. RPE signals in the dorsal and ventral striatum were scaled by the posterior probability of the hidden agent intervention, but in opposite directions. These findings are consistent with the Bayesian reinforcement learning model, which predicts that striatal learning is gated by causal inference.

Introduction

Beliefs about control over outcomes influence how people interact and learn from the world around them. For example, we recently demonstrated that by manipulating beliefs about causal structure, individuals showed asymmetric learning of positive and negative outcomes (Dorfman, Bhui, Hughes, & Gershman, 2019). Specifically, participants learned more from positive than from negative outcomes when hidden agents intervened adversarially, and conversely, learned more from negative than from positive outcomes when hidden agents intervened benevolently. A Bayesian model explained the complete pattern of asymmetries, and an extension of the model that inferred the probability of hidden agent intervention could capture performance in a more complex scenario in which the intervention probability was unknown. We have demonstrated that a Bayesian reinforcement learning process can explain biases produced by beliefs about hidden causes, but it remains unclear whether this process can be represented neurobiologically.

Converging evidence from prior studies has identified a suite of brain regions implicated in judgments of causal attributions and reinforcement learning. For example, inferior parietal regions have been linked to causal attributions of behavior and agency over actions (Farrer & Frith, 2002; Ruby & Decety, 2001). Further, there is some evidence suggesting that processes associated with external (e.g., attribution to another person) and internal agency (e.g. attribution to the self) may recruit distinct neural systems. For example, a meta-analysis examining the effects of agency beliefs on functional brain recruitment revealed that the recruitment of the superior temporal gyrus (STG), inferior parietal lobe (IPL), precuneus, pre-SMA, and dorsomedial prefrontal cortex (dmPFC) are selective to external agency beliefs, whereas activity in the bilateral insula, bilateral primary somatosensory cortex and left pre-motor cortex are associated with internal agency (Sperduti, Delaveau, Fossati, & Nadel, 2011). While external and

internal agency recruit distinct systems, there is also evidence for overlapping recruitment in regions such as the central gyrus, which suggests that these processes are not entirely distinct (Sperduti et al., 2011). In the learning literature, previous work in rodents, primates, and humans has shown that reward prediction errors (RPEs), which represent the difference between expected and received reinforcement, are encoded by dopaminergic neurons in the midbrain (Schultz, Dayan, & Montague, 1997), and prediction error information is projected to the striatum and frontal cortex (see (Niv, 2009) for a review). In human neuroimaging, RPE signal is found in regions of the striatum, prefrontal cortex (PFC) and orbitofrontal cortex (OFC) (McClure, Berns, & Montague, 2003; O'Doherty, Dayan, Friston, Critchley, & Dolan, 2003). Importantly, there is also evidence for magnitude fluctuations of RPE-related activation. For example, striatal activation tracks with the magnitude of prediction errors (Bermudez & Schultz, 2010) and striatal RPEs are sensitive to distributional properties of reward (Park et al., 2012). Together, this evidence suggests that it is reasonable to expect striatal BOLD activation in our study to mimic the magnitude asymmetries from our previous behavioral work.

Prior human neuroimaging research has examined functional brain recruitment associated with perceptions of agency and reinforcement learning in parallel, however, less is known about the neural systems underlying how perceptions of agency interact with the ability to learn from positive and negative outcomes. Although no human neuroimaging work to date explicitly tests this, some research implicates reward-related regions in this interaction. For example, while doing a motion prediction task, participants showed greater activation in the anterior insula, anterior cingulate cortex, and midbrain for trials where they believe they caused a loss, but diminished activation in the ventral putamen for these same trials (Späti et al., 2014). In addition, Bhanji & Delgado (2014) have shown that the striatum is sensitive to changes in controllability,

and others have found that caudate and nucleus accumbens are sensitive to rewards that are chosen rather than passively received (Zink, Pagnoni, Martin-Skurski, Chappelow, & Berns, 2004).

In the present study, we seek to investigate the neural mechanisms for agency-modulated learning. Here, we are interested in whether these learning asymmetries are mirrored in striatal BOLD activation, and if by using subjective attribution judgments and isolating components of our computational model, we can determine that agency-modulated learning is governed by distinct brain regions.

Method

Participants & Data Inclusion

Participants were recruited from the University of California at Riverside SONA study pool system. A total of 36 right-handed adults (ages 18-24, mean = 19.3, 14 males, 24 females) participated in the study. Individual runs were evaluated for excessive head movement (> 4mm), but no runs were ultimately excluded for excessive movement. Eleven participants were excluded from analyses due to unavoidable technical issues (n = 5), and three subjects had one run excluded from analyses due to unavoidable technical issues. To meet the accuracy threshold, participants must have chosen the stimulus option with the higher reward probability for at least 50% of the trials, which indicates above chance performance, and participants who did not meet this threshold were excluded from analysis (n = 6). This left a total of 25 participants for the majority of analyses. An additional two participants, although meeting accuracy criterion for the task, did not have enough variation in their choices to be properly modeled, so some analyses where conditions were further split include 23 participants.

Figure 2.1. fMRI Task Schematic. At the start of each run, participants are told which territory they are in ("Tycoon" or "Bandit"). Trial components consist of a choice between two stimuli (2.5 s), a fixation ISI (1.5-2.5 s), feedback (win or loss; 1.5 s), a fixation ISI (2-4 s), and an intervention guess (3 s). At the end of each run, a fixation IRI was presented for 10 s plus the residual amount of time from the stimulus choice and intervention guess events. ISI denotes inter-stimulus interval. IRI denotes inter-run interval.

Task Details

The task was presented using PsychoPy software version 1.85.2 (Peirce, 2007), and displayed on a screen visible through a mirror attached to the head coil. Behavioral responses were collected with a MRI-compatible button box, and all participants used the index and middle finger of their right (dominant) hand to make responses. Prior to doing the task in the scanner, participants received verbal instructions and completed a practice version of the task.

In this study, participants completed a reinforcement learning task in which they encountered multiple learning environments. This task was modified from our previous behavioral task for use in the scanner (Fig. 2.1; Dorfman et al., 2019). Participants were instructed to imagine that they were mining for gold in the Wild West. On each trial, participants had to choose between one of two different colored mines by using the button box. After making a decision (*choice period*: 2.5 s), an inter-stimulus fixation interval (*ISI*: jittered between 1.5 s-2.5 s) was displayed. Participants then viewed feedback of either a reward (gold) or loss (rocks) for 1.5 s. Each mine in a pair produced a reward with 70% or 30% probability of winning gold. Participants were instructed that each reward yielded a small amount of bonus money and each

loss resulted in a subtraction of bonus money. In actuality, all participants received a $5 bonus at the end of the experiment.

Participants completed four blocks of 30 trials (120 total trials) in different "mining territories." A single block was presented for each functional run. Participants were instructed that different agents frequent each territory: a bandit will steal gold from the mines and replace it with rocks (adversarial condition) and a tycoon will leave extra gold in the mines (benevolent condition). During the task, participants completed two blocks of each condition, which were interleaved in a pseudorandomized fashion. The agents intervened on 30% of the trials, and participants were told this proportion explicitly at the start of the task, though they did not know unambiguously whether the agent intervened on any particular trial. While the underlying reward distributions (i.e., absent intervention) for the mines were 70% or 30%, the hidden agent intervened on 30% of trials (or 9 out of 30 trials). For example, the benevolent intervention produces rewards on 9 out of 30 trials and the adversarial intervention produces losses on 9 out of 30 trials. After feedback and the presentation of a second ISI (2-4 s), participants were asked whether they believed the outcome they received was a result of hidden agent intervention (binary response of "Yes" or "No"; 3s) and made their selection using the button box. During the stimulus choice and intervention guess period, the events would end as soon as a button was pressed. Residual time from these events was added to the inter-run interval (IRI; 10 s plus residual) at the end of each run/block.

Computational Model

Previously, we demonstrated that a computational model that accounts for beliefs about agency can explain asymmetric updating of positive and negative outcomes (Dorfman et al., 2019). This

model allows for a participant to update their estimate of the reward probability for a particular action according to a reinforcement learning equation that integrates perceptions of causal inference. Specifically, the learning rate changes across trials depending on the posterior probability of the hidden agent intervention, as computed by Bayes' rule. Intuitively, when the posterior probability is high, the learning rate is low, and the participant should suppress learning about the reward probabilities when they believe that the outcome was generated by an external force. Our hypothesized model includes a parameter, ψ, which reflects a trial-by-trial estimate of the participant's belief about non-intervention, conditional on the past history of actions and rewards for the relevant task block. In other words, a larger ψ value indicates more control over outcomes, and a smaller ψ value indicates less control over outcomes. In one version of our hypothesized model, we fixed the prior probability of hidden agent intervention at 30% to replicate the instructions that participants received in the task, while another version initialized the prior probability of intervention as the mean of the participant's subjective intervention judgments. The former we will refer to as the *fixed Bayesian model*, and the latter the *empirical Bayesian model*. We hypothesized that one of the two Bayesian models would best fit the behavior of our participants when compared to a six-learning rate model that cannot account for causal inference. The six-learning rate model fits a different learning rate for each condition by valence combination (e.g., positive Benevolent, negative Benevolent, etc.), and therefore it makes no assumptions about asymmetry.

In order to investigate whether there is an anatomical distinction in the striatum for learning during trials when an individual believes they have agency compared to trials when they do not have agency in a model-based way, we configured subcomponents of the empirical Bayesian model in order to formalize three distinct value representations: a baseline

representation, calculated as RPE, an *agency representation*, calculated as $RPE * \psi$, and a *non-agency representation*, calculated as $RPE * (1 - \psi)$. Put simply, the baseline representation does not take into account beliefs about control, the agency representation scales the prediction error by the posterior over controllability (ψ), and the non-agency representation scales the prediction error by the posterior of uncontrollability (the inverse of ψ). We hypothesized that these separable representations would parametrically modulate different regions of the striatum during feedback. Specifically, we expected to see ventral striatal activation for all three representations but enhanced dorsal striatal activation for the agency representation.

fMRI Data Acquisition & Preprocessing

Imaging data were collected on a 3.0 T Magnetom Prisma MRI scanner with the vendor 32-channel head coil (Siemens Healthcare, Erlangen, Germany) at the University of California at Riverside Center for Advanced Neuroimaging. A T1-weighted high-resolution multi-echo magnetization-prepared rapid-acquisition gradient echo (ME-MPRAGE) anatomical scan (van der Kouwe et al., 2008) of the whole brain was acquired for each subject prior to any functional scanning (176 sagittal slices, voxel size = 1.0 x 1.0 x 1.0 mm, TR = 2530 ms, TE = 1.69 - 7.27 ms, TI = 1350 ms, flip angle = 7°, FOV = 256 mm). Functional images were acquired using a T2*-weighted echo-planar imaging (EPI) pulse sequence that employed multiband RF pulses and Simultaneous Multi-Slice (SMS) acquisition (Moeller et al., 2010; Feinberg et al., 2010; Xu et al., 2013). In total, four functional runs were collected for each subject, with each run corresponding to a single task block, two for each condition (84 interleaved axial-oblique slices per whole brain volume, voxel size = 1.5 x 1.5 x 1.5 mm, TR = 2000 ms, TE = 30 ms, flip angle = 80°, in-plane acceleration (GRAPPA) factor = 2, multi-band acceleration factor = 3, FOV =

204 mm). Functional slices were oriented to a 30° tilt towards coronal from AC-PC alignment. The SMS-EPI acquisitions used the CMRR-MB pulse sequence from the University of Minnesota.

Functional data were preprocessed and analyzed using SPM12 (Wellcome Department of Imaging Neuroscience, London, UK). Each functional scan was realigned to correct for small movements between scans, producing an aligned set of images and a mean image for each subject. The high-resolution T1-weighted ME-MPRAGE images were then co-registered to the mean realigned images and the gray matter was segmented out and normalized to the gray matter of a standard Montreal Neurological Institute (MNI) reference brain. The functional images were then normalized to the MNI template (resampled voxel size 2 mm 343 isotropic), spatially smoothed with an 8 mm full-width at half-maximum (FWHM) Gaussian kernel, high-pass filtered at 1/128 Hz, and corrected for temporal autocorrelations using a first-order autoregressive model.

fMRI Analysis

General linear models (GLMs) included impulse regressors that were convolved with the hemodynamic response function. Temporal onsets for stimulus choice, feedback, and intervention judgment were models as regressors of interest. All trial events besides the regressors of interest for a particular GLM were included as nuisance regressors, with the exception of ISIs, ITIs, and IRIs. Trial-by-trial parameters from the computational models were included in relevant GLMs as parametric modulators. A constant regressor for baseline activity and six-parameter motion regressor were also included in all GLMs.

All group-level reported results are *t*-maps that have been whole-brain corrected at a voxel-wise threshold of $p < 0.001$ and a cluster-corrected at $p < 0.05$, family-wise error (FWE). Region labels are based on the Harvard-Oxford Cortical and Subcortical Atlases and the SPM Automated Anatomical Labeling atlas (AAL2; (TzourioMazoyer:2002bi; Rolls, Joliot, & Tzourio-Mazoyer, 2015) and voxel-coordinates are reported in Montreal Neurological Institute (MNI) space.

Results

Behavior & Computational Model

To verify that participants made attribution judgments that made sense given the experimental manipulation, we looked at their hidden agent intervention judgments by outcome valence and condition. Replicating results from our previous paper, we found overall that participants were more likely to believe that negative outcomes were more likely to be caused by the hidden agent compared to positive outcomes, $F(31) = 14.88$, $p < 0.001$. There was also a significant condition by outcome valence interaction, $F(31) = 74.7$, $p < 0.0001$, where participants were more likely to believe that the hidden agent had intervened after negative compared to positive outcomes in the adversarial condition, and after positive compared to negative outcomes in the benevolent condition.

We compared the empirical Bayesian model (posterior over hidden causes is set at the individual's mean subjective intervention judgment) , the fixed Bayesian model (posterior over hidden causes is set at 30%), and the six-learning rate model (separate learning rates for each condition by valence combination) using random-effects Bayesian model selection (Rigoux, Stephan, Friston, & Daunizeau, 2014; Stephan, Penny, Daunizeau, Moran, & Friston, 2009).

This procedure treats each participant as a random draw from a population-level distribution over models, which it estimates from the sample of model evidence values for each model. We used the Laplace approximation of the log marginal likelihood to obtain the model evidence values. For our comparison metric, we report 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. We found that the empirical Bayesian model ($PXP = 0.97$) is superior to both the fixed Bayesian model ($PXP = 0.02$) and the six-learning rate model ($PXP < 0.01$), and demonstrates our predicted 4-way asymmetry between valence and condition, replicating our previous studies (see Supplemental Materials, Fig. S2.1). To further interrogate our model, we can also look at how the model corresponds to actual participant behavior. We found that learning rates were significantly lower for trials where participants believed that the hidden agent intervened, compared to trials where they believed that the hidden agent did not intervene, $t(24) = -5.48$, $p < 0.0001$, $d = -1.10$ (Supplemental Materials, Fig. S2.2). In addition, participants' judgments about intervention also showed a significant median point-biserial correlation with the intervention predicted by the model, $r = 0.424$, $p < 0.0001$ (Supplemental Materials, Fig. S2.3).

fMRI Results

Asymmetric learning

Since previous work provides evidence that the magnitude of striatal activation can shift in accordance with changes in reward magnitude and prediction error strength, we expected to see magnitude differences in striatal activation for contrasts comparing positive and negative outcomes (e.g. winning gold compared to receiving a rock). We also hypothesized that the

magnitude of striatal BOLD response would track the learning rate asymmetry across conditions shown in the task behavior.

The goal of these analyses was to identify neural regions that might exhibit asymmetry in activation magnitude for the interaction between condition and feedback type. We began with the simplest possible analysis to address this question by performing whole-brain analyses that assessed functional recruitment during feedback and compared recruitment that was enhanced for the benevolent relative to the adversarial conditions. Results revealed that there was greater activity in the left anterior cingulate and the left medial prefrontal cortex for the benevolent compared to adversarial condition during feedback (Table 2.1).

Table 2.1: benevolent > adversarial			MNI Coordinates		
Region Label	Extent	t-value	x	y	z
Paracingulate Gyrus	188	6.198	-12	38	-6
Frontal Pole	261	5.101	-16	58	4
Superior Frontal Gyrus	159	4.975	-18	36	50

Next, we computed analyses to identify functional recruitment that was enhanced for positive relative to negative feedback. These whole-brain analyses revealed significant differences in activation for win feedback compared to loss feedback (wins > losses) in, the striatum and prefrontal cortices (PFC), regions that have been consistently implicated in incremental reward learning (Table 2.2). We investigated these findings further by computing separate contrasts for feedback type (wins versus losses) in the different experimental conditions (benevolent, adversarial). In the benevolent condition (wins benevolent > losses benevolent), there was differential activation for wins relative to losses in the left hippocampus (149 voxels at X = -28, Y = -8, Z = -24) and the left medial prefrontal cortex (137 voxels at X = -2, Y = 62, Z = -8). In the adversarial condition, (wins adversarial > losses adversarial) there was differential

activation in the striatum and mPFC (Table 2.3). To identify whether functional recruitment of these regions exhibited asymmetric recruitment for positive and negative outcomes across conditions, we conducted a 4-way interaction. However, when we computed a 4-way interaction contrast of feedback type and condition type, we did not find the hypothesized asymmetric magnitude differences in the striatum. A single cluster of activation in the right medial orbitofrontal cortex (OFC) withstood FWE correction (153 voxels at X = 22, Y = 30, Z = -18).

Table 2.2: wins > losses			MNI Coordinates		
Region Label	Extent	t-value	x	y	z
Accumbens	5475	10.077	10	6	-10
Occipital Fusiform Gyrus	420	6.897	-36	-74	-32
Inferior Temporal Gyrus	646	6.561	-58	-54	-16
Cerebral White Matter	157	6.248	30	-42	28
Middle Temporal Gyrus	285	6.065	60	-6	-26
Occipital Pole	145	5.962	30	-92	-8
Cerebellum	173	5.233	18	-46	-20
Superior Frontal Gyrus	245	5.173	20	34	52
Inferior Temporal Gyrus	141	5.097	52	-42	-22
Superior Frontal Gyrus	336	5.078	-18	24	54
Putamen	231	5.050	-24	-8	18
Lateral Occipital Cortex	162	4.910	20	-76	42
Precuneus	322	4.644	-4	-60	34
Rectus	5475	3.675	2	54	-18

Table 2.3: wins adversarial – losses adversarial			MNI Coordinates		
Region Label	Extent	t-value	x	y	z
Accumbens	5844	12.530	10	6	-8
Thalamus	291	7.743	-18	-10	16
Precuneus	1404	6.955	-12	-56	36
Inferior Temporal Gyrus	1914	6.951	-56	-54	-16
Thalamus	364	6.866	12	-14	18

Lateral Occipital Cortex	164	6.093	58	-64	-14
Left Cerebral White Matter	135	6.069	-26	2	32
Middle Temporal Gyrus	284	5.784	62	-8	-28
Temporal Fusiform Cortex	202	5.556	-28	-46	-26
Temporal Fusiform Cortex	420	5.468	40	-36	-18
Frontal Pole	149	5.196	-20	44	36
Superior Frontal Gyrus	165	4.626	-26	18	62

Beliefs about hidden agent intervention

To interrogate the neural influence of beliefs about agency on learning, we contrasted wins and losses for trials where participants believed the hidden agent caused a given outcome in one model (wins for no agency trials > losses for no agency trials), and we contrasted wins and losses for trials where participants believed they caused a given outcome in another model (wins for agency trials > losses for agency trials). When participants believed the hidden agent had control over feedback, we found clusters with increased activity in the medial prefrontal cortex (mPFC) and the ventral striatum (bilateral nucleus accumbens), extending into the right amygdala for wins relative to losses (Table 2.4). Next, we examined the contrast of win > loss for trials in which participants believed that their actions produced the outcomes they received. Interestingly, the model looking at trials where participants believed they had agency over outcomes showed significantly increased activation in overlapping ventral striatal and mPFC regions, but also displayed activation extending into dorsal striatal regions, including the caudate, suggesting that RPEs are represented differentially in striatum based on subjective beliefs about agency (Table 2.5).

Table 2.4: wins/no agency > losses/no agency			MNI Coordinates		
Region Label	Extent	t-value	x	y	z
Frontal Pole	1158	7.030	-8	68	4
Left Putamen	210	6.158	-16	4	-10
Right Accumbens	199	5.409	12	8	-12
Superior Frontal Gyrus	262	5.098	-22	36	46

Table 2.5: wins/agency > losses/agency			MNI Coordinates		
Region Label	Extent	t-value	x	y	z
Right Caudate	1574	9.875	4	10	-6
Right Cerebral White Matter	319	8.520	18	28	-14
Left Precuneous Cortex	829	7.045	-8	-58	28
Left Precentral Gyrus	2101	6.457	-28	-20	62
Right Cerebral White Matter	222	6.257	22	4	24
Paracingulate Gyrus	290	5.903	10	52	6
Left Cerebellum	151	5.801	-36	-74	-34
Left Middle Temporal Gyrus	142	5.200	-66	-50	-4
Left Cerebral White Matter	251	5.102	-16	-16	32
Right Caudate	130	4.477	16	-14	22
Left Precentral Gyrus	2101	3.808	-20	-24	72

To investigate this hypothesis, we performed a model-based comparison using model-derived quantities as opposed to variables derived from behavior. To further interrogate learning mechanisms, we conducted region of interest analyses within the ventral and dorsal striatum. These regions were chosen because a large body of prior work has shown that activity in these regions is associated with RPE during learning. In the present study, we sought to compare how well three possible GLMs accounted for activation in the nucleus accumbens and caudate: 1) RPE alone, 2) $RPE * \psi$, and 3) $RPE * (1 - \psi)$. In order to compare dorsal and ventral striatal activation we generated anatomical masks of the bilateral nucleus accumbens and the bilateral

caudate (Pauli, Nili, & Tyszka, 2018). In order to compare the models, we approximated the log model evidence for each subject using the Bayesian information criterion, which we then used to compute the PXP for each model. This comparison showed that $RPE * (1 - \psi)$ (*PXP* = .4895) better accounted for activity in the nucleus accumbens compared to $RPE * \psi$ (*PXP* = .30) or RPE alone (*PXP* = .21). Activity in the bilateral caudate, however, was better accounted for by $RPE * \psi$ (*PXP* = .68) compared to $RPE * (1 - \psi)$ (*PXP* = .28) or the RPE-alone model (*PXP* = .04). These results suggest that ventral striatum encodes value regardless of agency, whereas dorsal striatum encodes actions over which the participant presumed to have agency.

To explore whether our task agency manipulation elicits differential brain activation in regions typically associated with agency judgments, we first compared activation for trials where participants believed that the hidden agent caused the outcome to trials where participants did not believe the hidden agent caused the outcome during the causal attribution phase (no agency > agency). This contrast revealed enhanced activation in left post-central gyrus (350 voxels at X = -40, Y = -18, Z = 48). However, no negative activation survived correction for this contrast, suggesting that no regions are preferentially active for beliefs about agency compared to beliefs about non-agency. We sought to verify this finding by identifying regions that parametrically tracked with ψ, so we computed a GLM that including a parametric modulator for the probabilistic belief about intervention at feedback (ψ). Our model parameter, ψ, represents the posterior over hidden causes, with higher values for ψ accounting for a stronger sense of agency. Using ψ as a single parametric modulator, a whole-brain analysis demonstrated significant activation in the right middle temporal gyrus (191 voxels at X = 54, Y = -26, Z = -12). Because higher values of ψ are equivalent to more agency, this analysis is theoretically equivalent to an

agency > non-agency contrast. However, using ψ as a parametric modulator as opposed to using participant's subjective beliefs, provides us with a continuous rather than a binary measure.

We also split the intervention beliefs by feedback type (wins and losses) and found significantly different activation in the left paracentral lobule when comparing trials where the participants believed the hidden agent caused a loss to trials where the participants believed they caused a loss (loss/no agency > loss/agency), modeled at the causal attribution phase (271 voxels at X = -12, Y = -26, Z = 64). However, we did not find any significant differences in activation for the contrast comparing win events during trials where the participants believed they did not have control to win events during trials where participants believed they did have control (wins/no agency > wins/agency). These findings suggest that beliefs about agency are particular salient for loss trials.

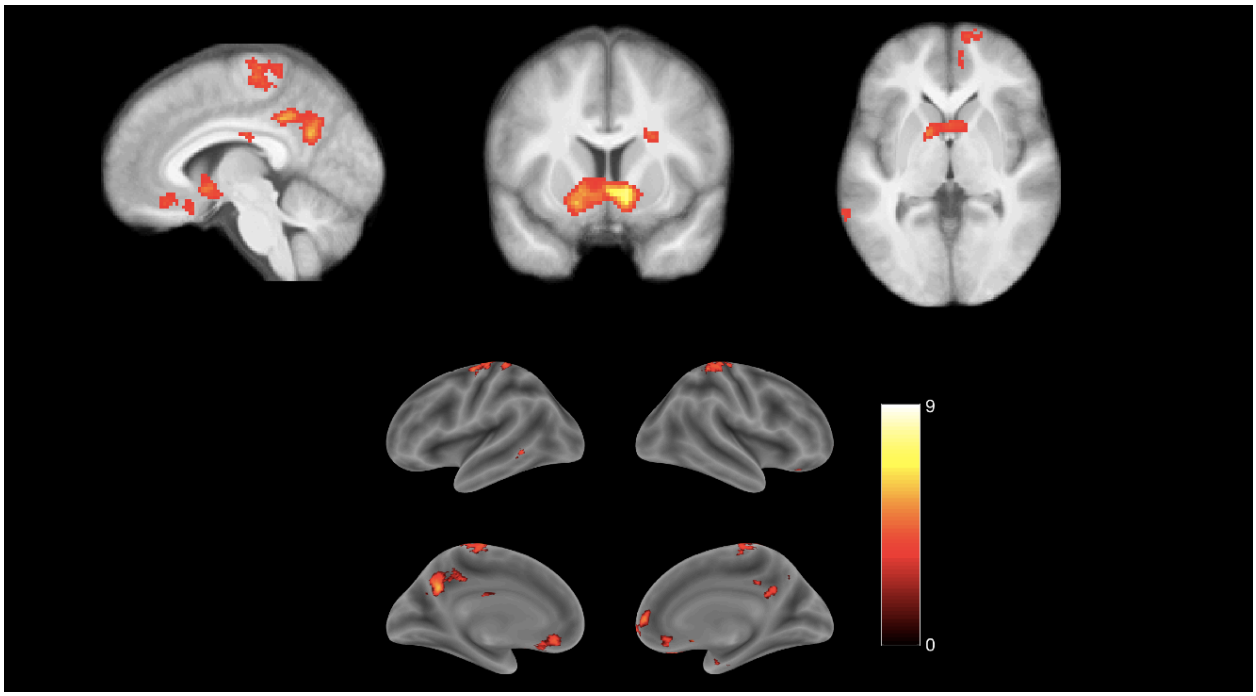


Figure 2.2. Brain areas activated by prediction errors modulated by the posterior over non-intervention, $RPE * \psi$; $p < 0.001$, whole-brain cluster FWE corrected at $p < 0.05$.

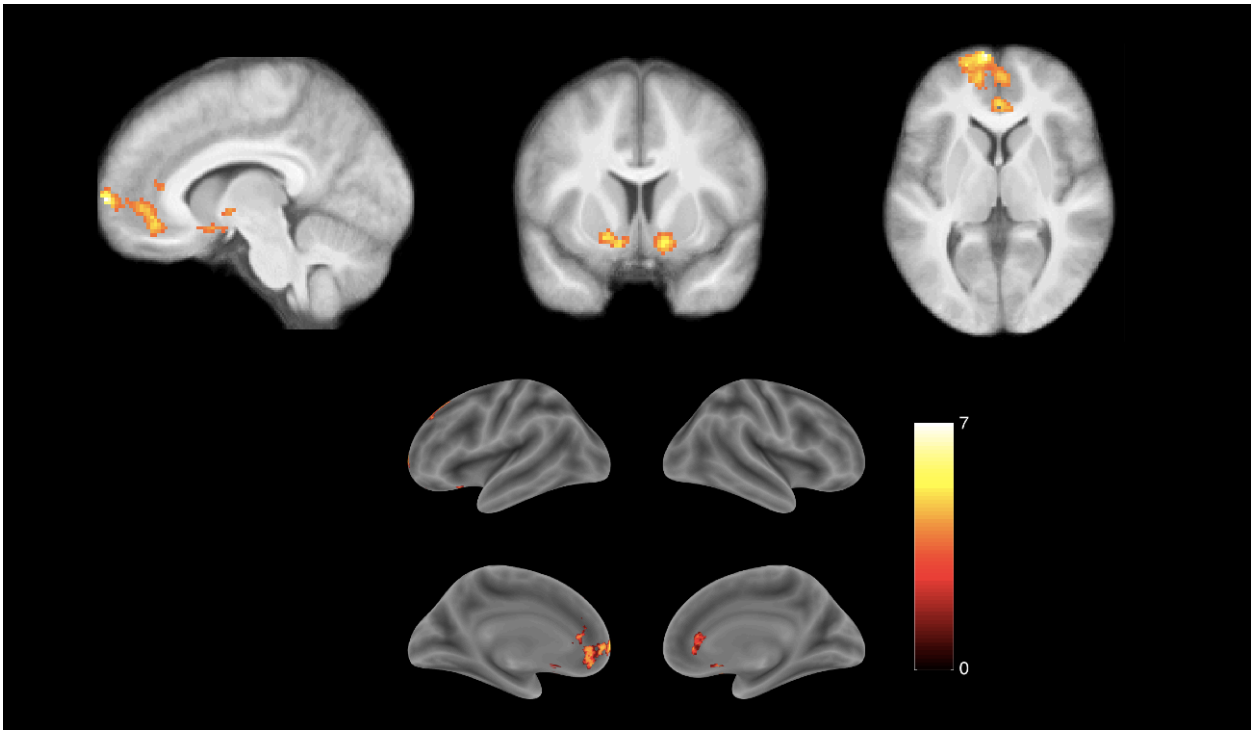


Figure 2.3. Brain areas activated by prediction errors modulated by the posterior over intervention, $RPE * (1 - \psi)$; $p < 0.001$, whole-brain cluster FWE corrected at $p < 0.05$.

Discussion

The present study sought to uncover the neurobiological mechanisms that modulate the interaction between learning from positive and negative feedback and beliefs about agency. We used functional neuroimaging to probe the brain regions that are associated with learning about probabilistic outcomes and agency judgments. Based on our previous work, we expected to see a difference in learning from positive versus negative outcomes that could be flexibly manipulated as beliefs about agency changed across task conditions. While we successfully replicated our behavioral findings, we did not see the hypothesized asymmetry in striatal prediction error signal. While surprising, some research has shown that striatal activation does not always track with behavioral biases in reward estimation. For example, behavioral reward probability distortions have been found to align with prefrontal activation, but not striatal activation. In the striatum, activation remains linear despite behavioral biases (Tobler, Christopoulos, O'Doherty, Dolan, & Schultz, 2008). Relatedly, Kool, Getz, and Botvinick showed that participants overestimate the probability of outcomes where they have control, but do not do so for outcomes that are out of their control, and that this asymmetry is not mirrored in striatal RPEs, as hypothesized (Kool, Getz, & Botvinick, 2013). We did, however, see activation in well-validated reward-related regions, such as mPFC and striatum when conducting basic win > loss contrasts. These results suggest that although striatum is representing RPEs generally, it does not seem to be representing behavioral biases in learning from positive and negative outcomes when encountering distinct agency manipulations.

Interestingly, when we used participants' subjective beliefs about control over outcomes, we found that more dorsal regions of the striatum were active when individuals believe they have more control. We then investigated this finding using a model-based analysis by comparing

activation in ventral (nucleus accumbens) and dorsal (caudate) anatomical ROIs with distinct subcomponents of our computational model. We found dorsal striatal activation was best explained by RPEs that were scaled by the parameter representing the posterior over hidden causes, ψ, but ventral striatal activation was best explained by RPEs that were scaled by 1- ψ. RPEs alone did not best explain activation in either region. We also explored the neural correlates for judgments about agency and found activation in the postcentral gyrus. Previous work has implicated parietal cortex as one of the regions important for agency judgments.

One method for exploring the relationship between controllability and reinforcement learning is by comparing instrumental and Pavlovian learning. These two distinct types of learning inherently include differences in controllability – with instrumental contexts involving outcome-linked, self-generated actions, and Pavlovian contexts involving involuntary associations between stimuli and responses. Studies have overwhelmingly implicated the dorsal striatum, specifically, the caudate nucleus, in instrumental learning tasks. Work by Tricomi and colleagues showed that dorsal striatal activity was associated with a button press contingent with participants' actions compared to button presses that were non-contingent (Tricomi, Delgado, & Fiez, 2004). One of the most direct tests of an agency-modulated dorsal-ventral striatal dissociation is work by O'Doherty and colleagues, where they reported ventral striatal activation for both Pavlovian and instrumental reward prediction errors (RPEs), and dorsal striatal (anterior cingulate) activation for instrumental RPE (O'Doherty et al., 2004). Our results suggest that as prediction errors are modulated by stronger beliefs of agency over outcomes, perhaps representing a form of enhanced instrumental learning, functional recruitment extends to more dorsal regions. However, it is important to note that the model comparisons are not decisive, especially for the ventral striatal activation, and when we discuss a distinction between dorsal

and ventral regions with regard to these results, we are noting a relative functional distinction and not a precise anatomical separation. One reason our model comparison results may not be decisive is because, in line with previous research, both instrumental and Pavlovian learning processes have been found to activate the ventral striatum, while instrumental processes more often distinctly activate the dorsal striatum. These results, in conjunction with the fact that we do not see asymmetric valence-tracking in the striatum, are in line with theories that propose that striatum encodes action tendency as opposed to valence (Guitart-Masip, Duzel, Dolan, & Dayan, 2014). However, this conclusion is quite speculative considering the minor evidence we provide, and future work should explore striatal activation during instrumental and Pavlovian learning while manipulating beliefs about agency within positive and negative outcomes.

Acknowledgements

We would like to thank Leah Somerville and Katie Insel for helpful discussion. Funding for this work was provided by the National Institutes of Health (CRCNS 1R01MH109177), the Office of Naval Research (N00014-17-1-2984) and the Alfred P. Sloan Foundation.

References

Bermudez, M. A., & Schultz, W. (2010). Responses of Amygdala Neurons to Positive Reward-Predicting Stimuli Depend on Background Reward (Contingency) Rather Than Stimulus-Reward Pairing (Contiguity). *Journal of Neurophysiology*, *103*(3), 1158–1170. http://doi.org/10.1152/jn.00933.2009

Dorfman, H. M., Bhui, R., Hughes, B. L., & Gershman, S. J. (2019). Causal Inference About Good and Bad Outcomes. *Psychological Science*, *133*, In Press. http://doi.org/10.1177/0956797619828724

Farrer, C., & Frith, C. D. (2002). Experiencing Oneself vs Another Person as Being the Cause of an Action: The Neural Correlates of the Experience of Agency. *NeuroImage*, *15*(3), 596–603. http://doi.org/10.1006/nimg.2001.1009

Gershman, S. J. (2016). Empirical priors for reinforcement learning models. *Journal of Mathematical Psychology*, *71*, 1–6. http://doi.org/10.1016/j.jmp.2016.01.006

Gershman, S. J., Pesaran, B., & Daw, N. D. (2009). Human reinforcement learning subdivides structured action spaces by learning effector-specific values. *The Journal of Neuroscience: the Official Journal of the Society for Neuroscience*, *29*(43), 13524–13531. http://doi.org/10.1523/JNEUROSCI.2469-09.2009

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

Kool, W., Getz, S. J., & Botvinick, M. M. (2013). Neural representation of reward probability: evidence from the illusion of control. *Journal of Cognitive Neuroscience*, *25*(6), 852–861. http://doi.org/10.1162/jocn_a_00369

Niv, Y. (2009). Reinforcement learning in the brain. *Journal of Mathematical Psychology*, *53*(3), 139–154. http://doi.org/10.1016/j.jmp.2008.12.005

O'Doherty, J., Dayan, P., Schultz, J., Deichmann, R., Friston, K., & Dolan, R. J. (2004). Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science (New York, NY)*, *304*(5669), 452–454. http://doi.org/10.1126/science.1094285

Park, S. Q., Kahnt, T., Talmi, D., Rieskamp, J., Dolan, R. J., & Heekeren, H. R. (2012). Adaptive coding of reward prediction errors is gated by striatal coupling. *Proceedings of the National Academy of Sciences*, *109*(11), 4285–4289. http://doi.org/10.1073/pnas.1119969109

Pauli, W. M., Nili, A. N., & Tyszka, J. M. (2018). A high-resolution probabilistic in vivo atlas of human subcortical brain nuclei. *Scientific Data*, *5*, 180063–13. http://doi.org/10.1038/sdata.2018.63

Rigoux, L., Stephan, K. E., Friston, K. J., & Daunizeau, J. (2014). Bayesian model selection for group studies - Revisited. *NeuroImage*, *84*(C), 971–985. http://doi.org/10.1016/j.neuroimage.2013.08.065

Rolls, E. T., Joliot, M., & Tzourio-Mazoyer, N. (2015). Implementation of a new parcellation of the orbitofrontal cortex in the automated anatomical labeling atlas. *NeuroImage*, *122*(C), 1–5. http://doi.org/10.1016/j.neuroimage.2015.07.075

Ruby, P., & Decety, J. (2001). Effect of subjective perspective taking during simulation of action: a PET investigation of agency. *Nature Neuroscience*, *4*(5), 546–550. http://doi.org/10.1038/87510

Schultz, W., Dayan, P., & Montague, P. R. (1997). A Neural Substrate of Prediction and Reward. *Science (New York, NY)*.

Späti, J., Chumbley, J., Brakowski, J., Dörig, N., Grosse Holtforth, M., Seifritz, E., & Spinelli, S. (2014). Functional lateralization of the anterior insula during feedback processing. *Human Brain Mapping*, *35*(9), 4428–4439. http://doi.org/10.1002/hbm.22484

Sperduti, M., Delaveau, P., Fossati, P., & Nadel, J. (2011). Different brain structures related to self- and external-agency attribution: a brief review and meta-analysis. *Brain Structure & Function*, *216*(2), 151–157. http://doi.org/10.1007/s00429-010-0298-1

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

Tobler, P. N., Christopoulos, G. I., O'Doherty, J. P., Dolan, R. J., & Schultz, W. (2008). Neuronal Distortions of Reward Probability without Choice. *Journal of Neuroscience*, *28*(45), 11703–11711. http://doi.org/10.1523/JNEUROSCI.2870-08.2008

Tricomi, E. M., Delgado, M. R., & Fiez, J. A. (2004). Modulation of caudate activity by action contingency. *Neuron*, *41*(2), 281–292.

Zink, C. F., Pagnoni, G., Martin-Skurski, M. E., Chappelow, J. C., & Berns, G. S. (2004). Human striatal responses to monetary reward depend on saliency. *Neuron*, *42*(3), 509–517.

Paper 3: Bayesian Arbitration Between Pavlovian and Instrumental Control

Hayley M. Dorfman and Samuel J. Gershman

(In Preparation)

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.

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 approach 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. The instrumental predictor thus 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 (Figure 3.1, 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 (its bets 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 (Figure 3.1, 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).

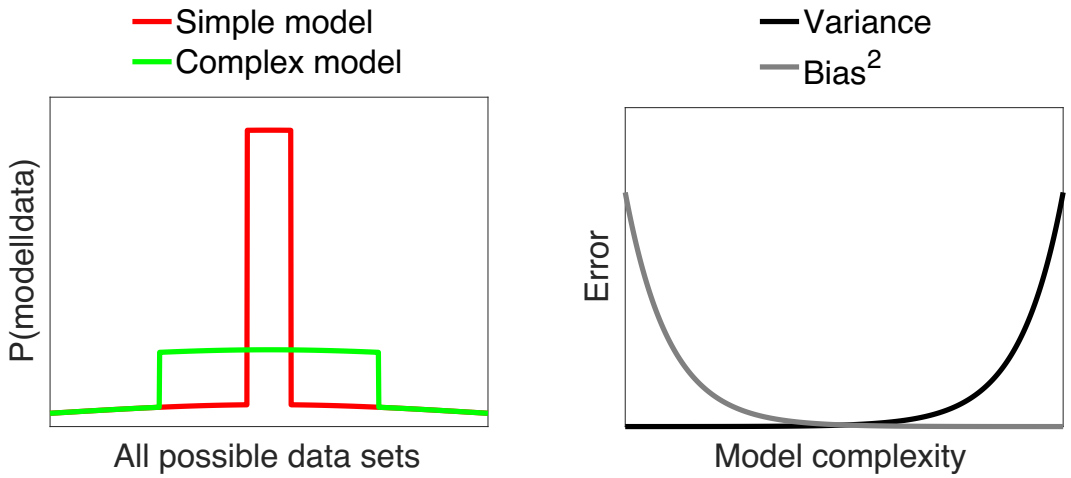


Figure 3.1. Two perspectives on model complexity. (**left**) Bayesian Occam's razor. Complex models distribute their probability across many different data sets, and thus, get less credit for observing any particular data set,

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.

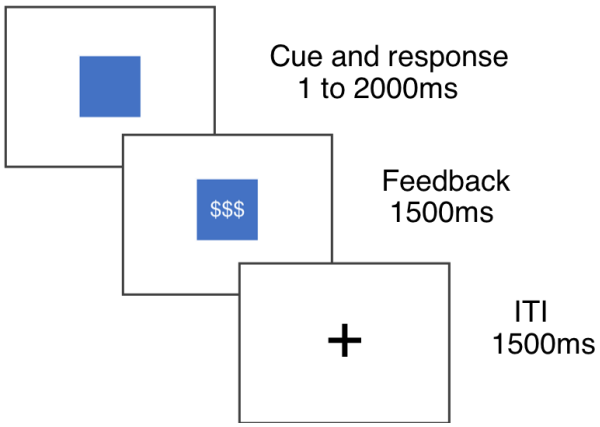
Applying these ideas to arbitration between Pavlovian and instrumental control, a key determinant of the optimal model complexity is *controllability of reward* (Huys & Dayan, 2009; Moscarello & Hartley, 2017). If rewards are uncontrollable (actions do not affect reward rate), then the simpler Pavlovian 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 (Figure 3.2). Experiment 1 manipulated reward controllability between participants, whereas Experiment 2 manipulated it within participants.

a



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

Figure 3.2. 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: $N = 212$). 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 (Figure 3.2) but differ in whether *w* is set

exogenously (in the case of the fixed model) or endogenously (in the case of the adaptive model).

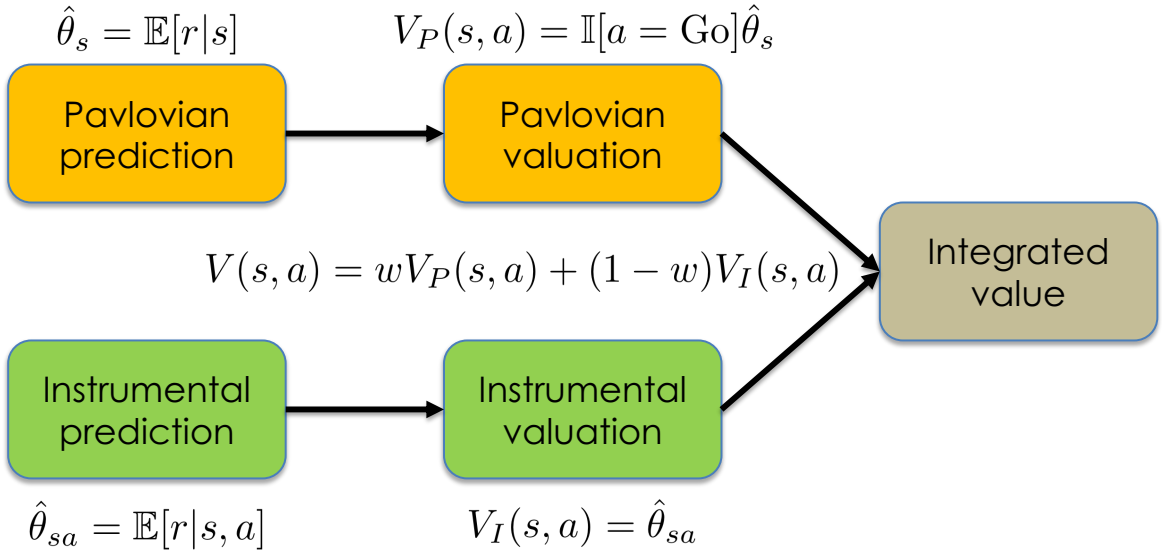


Figure 3.2. Information processing architecture. Pavlovian and instrumental prediction and valuation combine into a single value. This integrated value includes a weighting parameter (w) that represents the evidence for the uncontrollable environment (i.e., in favor of the Pavlovian predictor).

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. 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.

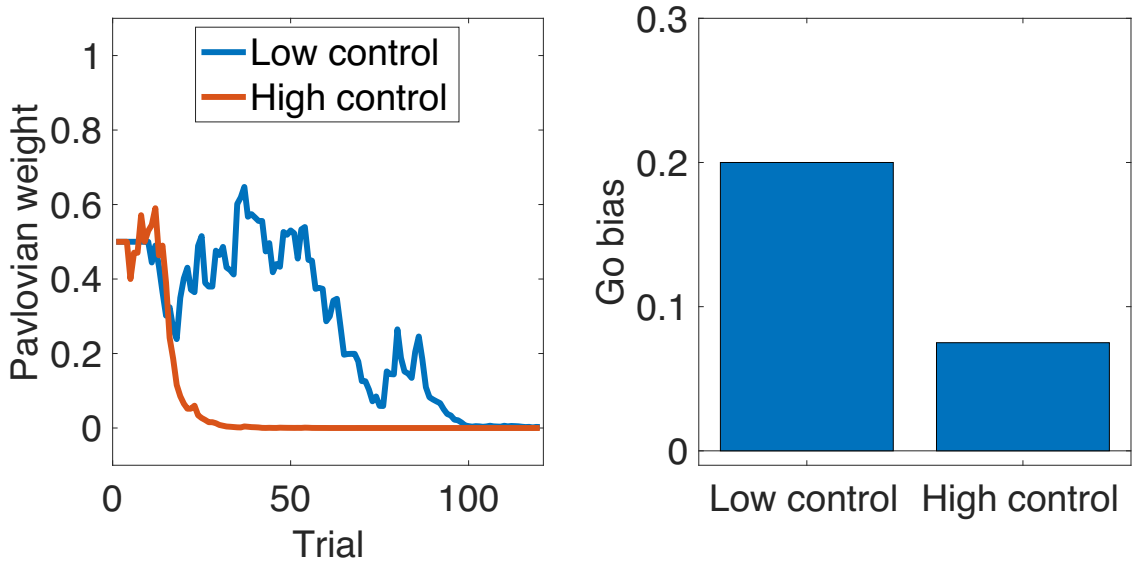


Figure 3.3. Model Simulations. Our model demonstrates a greater reliance on the Pavlovian system in the Low Control condition compared to the High Control condition (**a**) Across all trial types, the Pavlovian weight derived from the model is greater for the Low Control condition. (**b**) The model 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 3.3 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. 3.4). 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 et al., 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. 3.4), 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. 3.3). 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. 3.5). 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. 3.6). 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: $p < 0.001$).

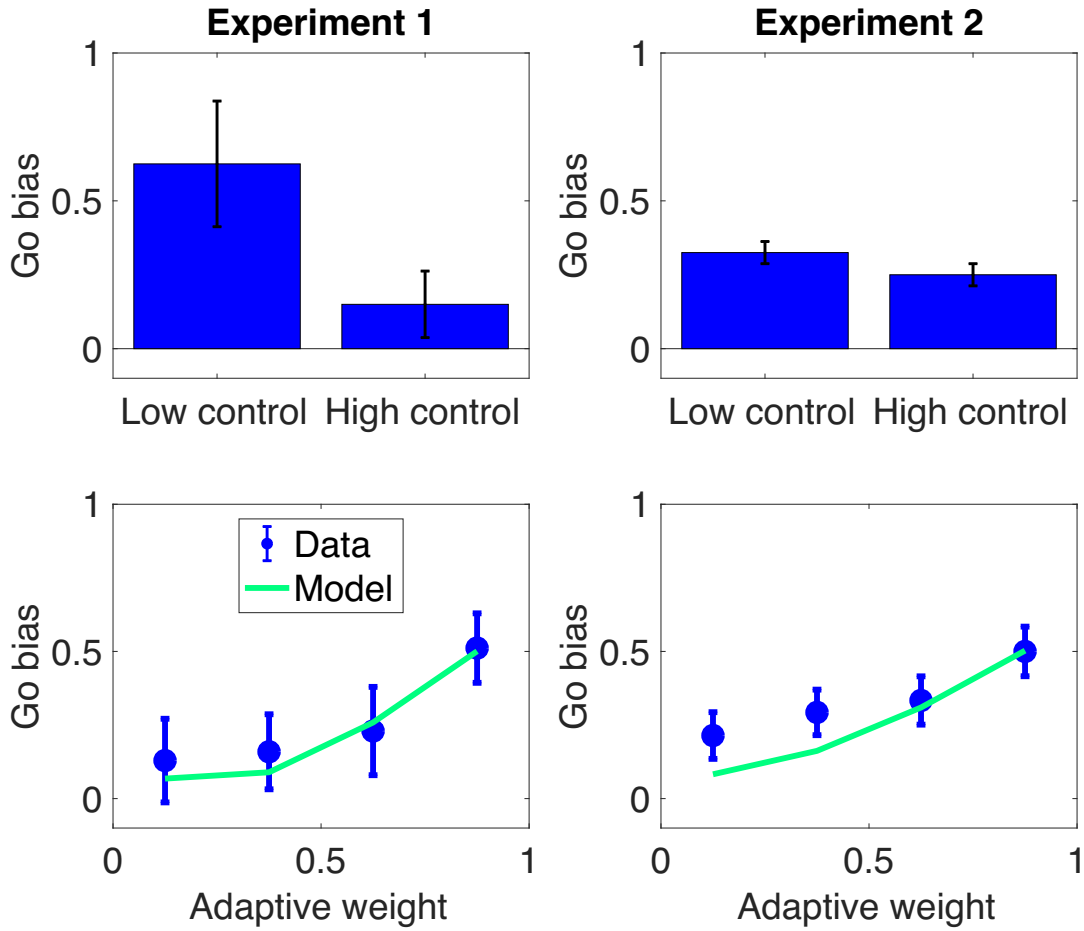


Figure 3.4. (**Top**) Go bias for low and high control conditions in Experiment 1 (**left**) and Experiment 2 (**right**). (**Bottom**) The adaptive model captures within-participant variability in Go bias, plotted as a function of Pavlovian weight (w) quantile for Experiment 1 (**left**) and Experiment 2 (**right**). Error bars show bootstrapped 95% confidence intervals around the median.

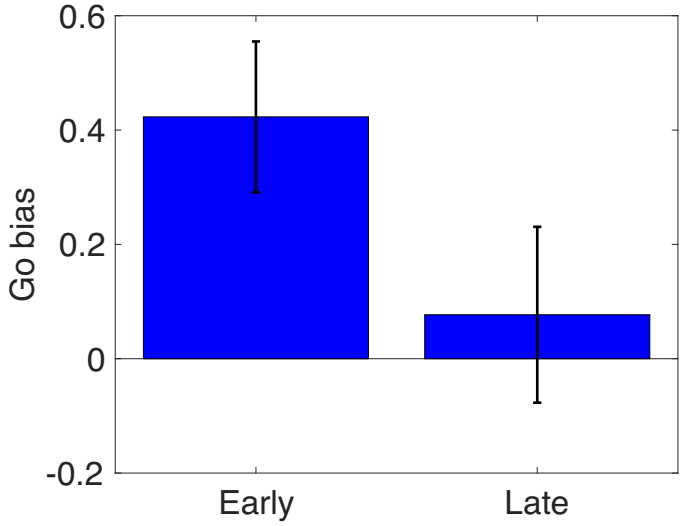


Figure 3.5. Go bias for early and late trials. Go bias in Experiment 1 is larger on the first 40 trials compared to the last 40 trials. Error bars show bootstrapped 95% confidence intervals around the median.

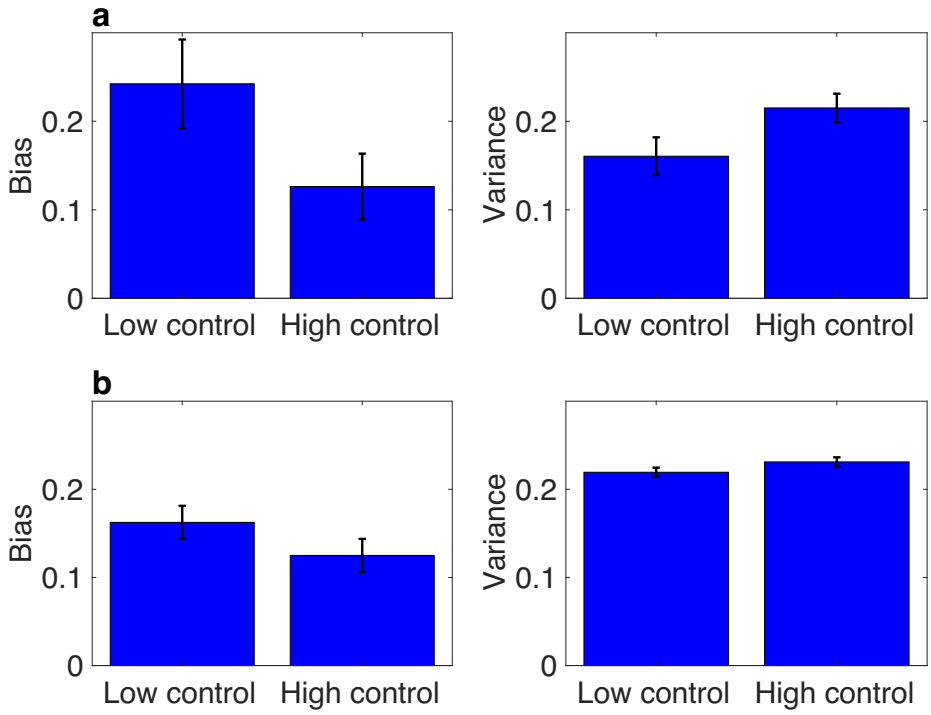


Figure 3.6. Bias and variance of choice behavior for Experiment 1 (**top**) and Experiment 2 (**bottom**). Error bars show bootstrapped 95% confidence intervals around the median.

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.

Acknowledgements

The authors would like to thank Rebecca Hao for her help with the initial setup for this study. Funding for this work was provided by the National Institutes of Health (CRCNS 1R01MH109177), the Office of Naval Research (N00014-17-1-2984) and the Alfred P. Sloan Foundation.

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

Lieder, F., & Goodman, N. D. (2013). Learned helplessness and generalization. *Cognitive Science*.

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

General Discussion

Summary of findings

In a series of three papers, this dissertation examined the behavioral and neurobiological processes that contribute to causal inference and utilized a modern computational framework to understand how these inferences modulate reinforcement learning. **Paper 1** investigated to what extent causal attributions to internal or external forces influence how individuals learn from positive and negative outcomes and the underlying computations that give rise to learning in this context. Across two experiments, results demonstrated that overall, individuals learned more from positive compared to negative feedback. This effect was modulated by agency, such that learning was magnified when participants believed they had more control over their outcomes. Learning could be magnified or dampened depending on whether the participant was in a condition that allotted them more or less agency over outcomes. A novel Bayesian reinforcement learning model that incorporated estimates of causal beliefs could predict the asymmetric learning by valence interaction that was demonstrated in the behavior. Participants also showed a self-preservation bias, where they were more likely to believe that they were responsible for positive outcomes and the hidden agent was responsible for negative outcomes.

Paper 2 explored how beliefs about agency and the associated impact on feedback-based learning asymmetries are represented in the brain. The valence-dependent learning asymmetries found in Paper 1 were replicated in Paper 2, and the same computational model was able to predict participant behavior. Paper 2 did not show that striatal prediction errors scaled along with causal inference, as hypothesized. However, we did find evidence for partial functional separation within the striatum for learning that was modulated by beliefs about agency.

Specifically, the ventral striatum tracked reward prediction errors, regardless of beliefs about control, but dorsal regions tracked RPEs that were scaled by beliefs about control.

In order to further examine the precise learning processes that are influenced by outcome controllability and to integrate our previous work with the broader reinforcement learning literature, **Paper 3** sought to provide a computational explanation for the arbitration between Pavlovian and instrumental learning. By manipulating the controllability of rewarding outcomes, we showed that Pavlovian processes were favored when controllability was low, and instrumental processes were favored when controllability was high, suggesting that balance between these two learning systems is reliant on an assessment of model complexity.

Implications

A cohesive framework

This work sought to integrate classic discoveries from clinical, social, and experimental psychology into a unified reinforcement learning framework with direct neurobiological correlates. Previous work on asymmetric evaluation of positive and negative outcomes has been notably mixed. It has been shown that both humans and animals learn faster from punishments compared to rewards and pay more attention to negative events (Baumeister et al., 2001; Taylor, 1991). Nonetheless, reinforcement learning studies have demonstrated higher learning rates for both positive (Kuzmanovic et al., 2016; Lefebvre et al., 2017; Moutsiana et al., 2015) and negative (Christakou et al., 2013; Gershman, 2015b; Niv et al., 2012) prediction errors. By providing a mechanistic explanation for how positive and negative outcomes might be favored in different contexts, as modulated by beliefs about agency, we can help explain these discrepancies in the literature.

While one of the benefits to using mathematical models is that they produce quantifiable parameters that can be used to investigate individual or group differences, perhaps the greatest benefit is that computational models formalize a hypothesis for a latent process such as agency. Successful models get us closer to uncovering how precisely these latent processes come about, which is particularly useful when studying behavior that is difficult or impossible to measure via traditional methods.

Clinical implications

Differences in learning from valenced feedback are important for understanding a variety of cognitive biases and psychopathological outcomes. Previous work has shown that biases in valenced learning are associated with depression, anhedonia, and pessimistic attitudes. For example, patients with depression and anhedonia exhibit blunted learning for both rewards and punishments (Chase et al., 2010), and depressed participants accurately recall negative outcomes while healthy participants underestimate the frequency of negative outcomes (Nelson & Craighead, 1977). In addition, research in both humans and animals has shown that beliefs about agency influence future instrumental responding. Depression and anxiety have also been linked to concepts such as *learned helplessness*, where exposure to uncontrollable punishments impairs future instrumental learning (Maier & Seligman, 1976; Mineka & Hendersen, 1985). In a seminal study by Alloy and Abramson, healthy participants exhibited an agency bias for desired outcomes, and a "non-agency" bias for undesired outcomes, while depressed participants showed no such bias (Alloy & Abramson, 1979). This work suggests that biased beliefs of control may be protective against depression, and that these cognitive distortions arise not solely from a belief that individuals have control over positive outcomes, but also that negative outcomes can be attributed

to someone or something outside of oneself. Consistent with this literature, our collective findings suggest that a latent factor, such as agency inference, may be mediating imbalanced learning for positive and negative outcomes. Since disruption of this imbalanced learning is a feature of disorders such as depression and anxiety, future work should examine the role of agency inference in these disorders.

Limitations

While one of the strengths of this work is that it proposes a cohesive framework for understanding the interaction between agency and learning, it will be imperative to investigate the limits of this framework in additional contexts. This is especially important given that this work has potentially useful implications for clinical outcomes. Since the emergence and presentation of psychopathologies are multifaceted, it will be necessary to test our framework in more complex scenarios in order to make strong claims. The work presented here is merely a first step toward answering important questions about learning processes, and all three papers make way for a number of new questions to be answered. The experiments we present are simple and controlled in order to act as proof of concepts, but this factor limits the number of questions we can answer and does not always provide a perfect analogy to real-world circumstances.

Conclusions

Extensive work suggests that beliefs about agency have substantial effects on how individuals learn from different types of outcomes. By utilizing computational models, neuroimaging, and flexible behavioral tasks, this dissertation investigated the behavioral and neurobiological pathways through which humans make inferences about hidden information and determined how

these inferences influence learning processes. We first found that inference about hidden agents modulates biased learning of positive and negative outcomes, demonstrating a mechanistic framework for understanding how causal attributions contribute to learning. Next, we showed that RPE signals in the dorsal and ventral striatum were scaled by both subjective and model-derived beliefs about agency, but in opposite directions, providing preliminary evidence that striatal learning is gated by causal inference. Finally, we show that controllability arbitrates the use of a Pavlovian over an instrumental learning system. Together, these results suggest that beliefs about agency are at least one factor that influences how we learn from feedback and how we decide the types of learning processes to utilize. Continuing this line of work will be particularly useful for elucidating how biased learning, and ultimately, outcomes like anxiety and depression, come about.

References

Alloy, L. B., & Abramson, L. Y. (1979). Judgment of contingency in depressed and nondepressed students: Sadder but wiser? *Journal of Experimental Psychology. General*, *108*(4), 441–485. http://doi.org/10.1037/0096-3445.108.4.441

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology*, *5*(4), 323–370. http://doi.org/10.1037//1089-2680.5.4.323

Chase, H. W., Frank, M. J., Michael, A., Bullmore, E. T., Sahakian, B. J., & Robbins, T. W. (2010). Approach and avoidance learning in patients with major depression and healthy controls: relation to anhedonia. *Psychological Medicine*, *40*(3), 433–440. http://doi.org/10.1017/S0033291709990468

Christakou, A., Gershman, S. J., Niv, Y., Simmons, A., Brammer, M., & Rubia, K. (2013). Neural and Psychological Maturation of Decision-making in Adolescence and Young Adulthood. *Journal of Cognitive Neuroscience*, *25*(11), 1807–1823. http://doi.org/10.1162/jocn_a_00447

Gershman, S. J. (2015). Do learning rates adapt to the distribution of rewards? *Psychonomic Bulletin & Review*, *22*(5), 1320–1327. http://doi.org/10.3758/s13423-014-0790-3

Kuzmanovic, B., Jefferson, A., & Vogeley, K. (2016). The role of the neural reward circuitry in self-referential optimistic belief updates. *NeuroImage*, *133*, 151–162.

Lefebvre, G., Lebreton, M., Meyniel, F., Bourgeois-Gironde, S., & Palminteri, S. (2017). Behavioural and neural characterization of optimistic reinforcement learning. *Nature Human Behaviour*, *1*, 0067. http://doi.org/10.1038/s41562-017-0067

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

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

Moutsiana, C., Charpentier, C. J., Garrett, N., Cohen, M. X., & Sharot, T. (2015). Human Frontal-Subcortical Circuit and Asymmetric Belief Updating. *The Journal of Neuroscience : the Official Journal of the Society for Neuroscience*, *35*(42), 14077–14085. http://doi.org/10.1523/JNEUROSCI.1120-15.2015

Nelson, R. E., & Craighead, W. E. (1977). Selective recall of positive and negative feedback, self-control behaviors, and depression. *Journal of Abnormal Psychology*, *86*(4), 379–388.

Niv, Y., Edlund, J. A., Dayan, P., & O'Doherty, J. P. (2012). Neural Prediction Errors Reveal a Risk-Sensitive Reinforcement-Learning Process in the Human Brain. *Journal of Neuroscience*, *32*(2), 551–562. http://doi.org/10.1523/JNEUROSCI.5498-10.2012

Taylor, S. E. (1991). Asymmetrical effects of positive and negative events: the mobilization-minimization hypothesis. *Psychological Bulletin*, *110*(1), 67–85.

Appendix

Supplemental Materials: Paper 1

Computational Models

Model Fitting. We used a softmax function to model choice probabilities, including a response stochasticity (inverse temperature) parameter and a "stickiness" parameter to capture choice autocorrelation (Gershman, Pesaran, & Daw, 2009). The Bayesian models were fit using maximum *a posteriori* estimation with empirical priors based on previous research (Gershman, 2016). Specifically, the prior distribution for the inverse temperature was $\beta \sim \text{Gamma}(4.82, 0.88)$, and for the stickiness parameter was $\rho \sim \mathcal{N}(0.15, 1.42)$.

Model Comparison. We used random-effects Bayesian model selection (Stephan, Penny, Daunizeau, Moran, & Friston, 2009) to compare models. This procedure treats each participant as a random draw from a population-level distribution over models, which it estimates from the sample of model evidence values for each model. We used the Laplace approximation of the log marginal likelihood to obtain the model evidence values. For our model comparison metric, we report the "protected exceedance probability" (*PXP*), the probability that a particular model is more frequent in the population than all other models under consideration. This is differentiated from an "exceedance probability" in that it considers the possibility that some differences in model evidence are due to chance.

Experiment 1:

Descriptive Model. In order to characterize learning rate asymmetries without committing to the assumptions of the Bayesian model, we fit a "descriptive" reinforcement learning model which updates reward probability estimates according to $\theta_{t+1} = \theta_t + \alpha_t(r_t - \theta_t)$, with separate learning rates for each combination of positive/negative outcome and the three experimental conditions (benevolent, adversarial, random).

Bayesian Reinforcement Learning Model. This model considers a problem in which a decision-maker must choose an option $c \in \{1, \ldots, C\}$ based on experienced rewards. Each option is associated with an intrinsic reward distribution, but a latent cause can intervene (denoted $z = 1$) to generate a different reward distribution. With probability $P(z = 0)$, the decision-maker receives a reward from the intrinsic distribution, $P(r|c, z = 0)$, or with probability $P(z = 1)$, she receives a reward determined by the latent cause, $P(r|z = 1)$. The marginal distribution over reward given the decision maker's choice is thus

(1)

$$P(r|c) = P(r|c, z = 0)P(z = 0) + P(r|z = 1)P(r|z = 1).$$

In our setting, rewards are binary and distributed according to a Bernoulli distribution for each option, $Bern(\theta^c)$. Thus, absent intervention, a reward of 1 is drawn with probability θ^c, and a reward of 0 is drawn with probability $1 - \theta^c$.

For our experimental paradigm, we define the latent cause as another agent (the "latent agent") that can allot outcomes for the decision maker. We define 3 different latent agent types:

• Benevolent latent agent: produces a reward regardless of the decision maker's choice. Formally, $P(r = 1|z = 1) = 1$.

• Adversarial latent agent: produces no reward regardless of the decision maker's choice. Formally, $P(r = 0|z = 1) = 1$.

• Random latent agent: produces reward with probability 0.5 regardless of the decision maker's choice. Formally, $P(r = 1|z = 1) = 1/2$.

The decision maker does not know the true reward probabilities of her options. She has a prior belief that the unknown parameters θ^c are independently distributed according to $Beta(a, b)$, which are then updated from experience. The Beta distribution parameters can be fit to choice data though for simplicity we assume both are equal to 1 in the estimation procedure, corresponding to a uniform distribution.

After choosing an action c_t and observing reward r_t on trial t, the decision maker updates her estimate of the reward probability θ^c according to a reinforcement learning equation that incorporates inference over latent causes. With a Beta prior and a Bernoulli reward distribution, the Bayesian update rule takes the form $\theta_{t+1} = \theta_t + \alpha_t(r_t - \theta_t)$, where α_t is a parameter representing the learning rate that scales the reward prediction error. This learning rate is based rationally on beliefs about the outcome's two possible sources: the action's intrinsic reward and the latent agent's intervention. These beliefs jointly determine the extent to which the participant attributes feedback to each source. The learning rate is given by

(2)

$$\alpha_t = \frac{P(z_t = 0|r_t, c_t)}{N_t^c + a + b}$$

where $N_t^c \approx \sum_{\tau \in \{1,\ldots,t|c_\tau = c\}} P(z_\tau = 0|r_\tau, c_\tau)$ is the sum of past beliefs about latent agent non-intervention on trials when the same option was chosen. The denominator reflects the magnitude of evidence about the intrinsic reward probability accumulated up to trial t (it is approximate because we assume for tractability that evidence provided by past feedback is not revised according to later information).

The learning rate's numerator is central to our present analysis: it encodes the degree to which feedback should be attributed to the intrinsic reward distribution rather than to the latent agent, and modulates the degree of learning based on whether feedback was positive or negative. The value of $P(z = 0|r, c)$ is stipulated by Bayes' rule:

(3)

$$P(z|r, c) = \frac{P(r|z, c)P(z)}{\sum_{z'} P(r|z', c)P(z')}.$$

This yields the following expressions which vary based on the combination of feedback and agent type:

Benevolent agent (negative feedback):

$$P(z = 0|r = 0, c) = 1$$

Benevolent agent (positive feedback):

$$P(z = 0|r = 1, c) = \frac{\theta^c P(z = 0)}{\theta^c P(z = 0) + P(z = 1)}$$

Adversarial agent (negative feedback):

$$P(z = 0|r = 0, c) = \frac{(1 - \theta^c) P(z = 0)}{(1 - \theta^c) P(z = 0) + P(z = 1)}$$

Adversarial agent (positive feedback):

$$P(z = 0|r = 1, c) = 1$$

Random agent (negative feedback):

$$P(z = 0|r = 0, c) = \frac{(1 - \theta^c) P(z = 0)}{(1 - \theta^c) P(z = 0) + P(z = 1)/2}$$

Random agent (positive feedback):

$$P(z = 0|r = 1, c) = \frac{\theta^c P(z = 0)}{\theta^c P(z = 0) + P(z = 1)/2}$$

where the probability of intervention $P(z = 1)$ is known. Subtly, the decision maker's inference about the latent agent's intervention depends on her existing estimate of θ^c.

The learning rate exhibits asymmetries depending on whether the latent agent tends to produce positive or negative outcomes. For example, when the agent is adversarial, positive outcomes can only come from the action itself, whereas negative outcomes are partly attributable to the external agent. Consequently, negative outcomes are less informative about the action's reward probability, corresponding to a lower learning rate.

	Parameter	Prior Distribution	Bounds	μ (mean)	95% CI
Bayesian RL Model	β (inverse temperature)	~ Gamma(4.82, 0.88)	[0.001, 20]	4.32	[3.71, 4.48]
	ρ (stickiness)	~ $\mathcal{N}$(0.15, 1.42)	[-5, 5]	1.51	[1.31, 1.72]

Figure S1.1. Computational model parameters for the best-fitting model in Experiment 1.

Experiment 2:

In Experiment 2, we consider the case where the decision-maker does not know the probability of intervention. We explored three possible models for this scenario: (1) the "adaptive Bayesian" model, which estimates the intervention probability directly, (2) the "fixed Bayesian" model, which treats the intervention probability as a free parameter, and (3) the "empirical Bayesian" model, which derives the intervention probability by averaging the participants' binary intervention judgments. The empirical Bayesian model was the best fitting model for our data across two independent samples (pooled and unpooled), with a *PXP* > 0.999.

Adaptive Bayesian Reinforcement Learning Model. For ease of comprehension, we define a new variable, ω, to represent the decision maker's estimate of the latent agent's intervention probability, $P(z_t = 1)$. This can be approximated by the average of past beliefs about

intervention, implying that the decision maker updates ω on each trial using:

(1)

$$\omega_{t+1} = \omega_t + \frac{1}{t + a + b} (P(z_t = 1|r_t, c_t) - \omega_t)$$

The ω update rule can be coupled with the θ update rule above by plugging ω_{t+1} into instances of $P(z_t = 1)$.

Fixed Bayesian Reinforcement Learning Model. Here, we fit the estimate of the latent agent's intervention probability, $P(z_t = 1)$, as a free parameter for each participant. This is integrated into the Bayesian Reinforcement Learning Model described above.

Empirical Bayesian Reinforcement Learning Model. This model calculates the decision-maker's average intervention judgment and utilizes this value for the estimate of the intervention probability $P(z_t = 1)$, providing an individualized estimate of intervention for each participant. This is integrated into the Bayesian Reinforcement Learning Model described above.

	Parameter	Prior Distribution	Bounds	μ (mean)	95% CI
Empirical Bayesian RL Model	β (inverse temperature)	$\sim$ Gamma(4.82, 0.88)	[0.001, 20]	4.10	[3.66, 4.05]
	Stickiness	$\sim \mathcal{N}(0.15, 1.42)$	[-5, 5]	1.22	[1.12, 1.32]

Table. S1.1. Computational model parameters for the best-fitting model ("empirical Bayesian") in Experiment 2. Prior Distribution: This model was fit using a maximum a posteriori estimation with empirical priors based on previous research (Gershman, 2016). Bounds: Limits set for the optimization procedure for each parameter. Mean: Mean of each parameter across all participants. CI: 95% confidence intervals for the parameters across all participants.

Behavioral Analyses

Win-stay Lose-shift. In order to explore participants' choices in a "model-free" way, we visualized their win-stay lose-shift behavior.

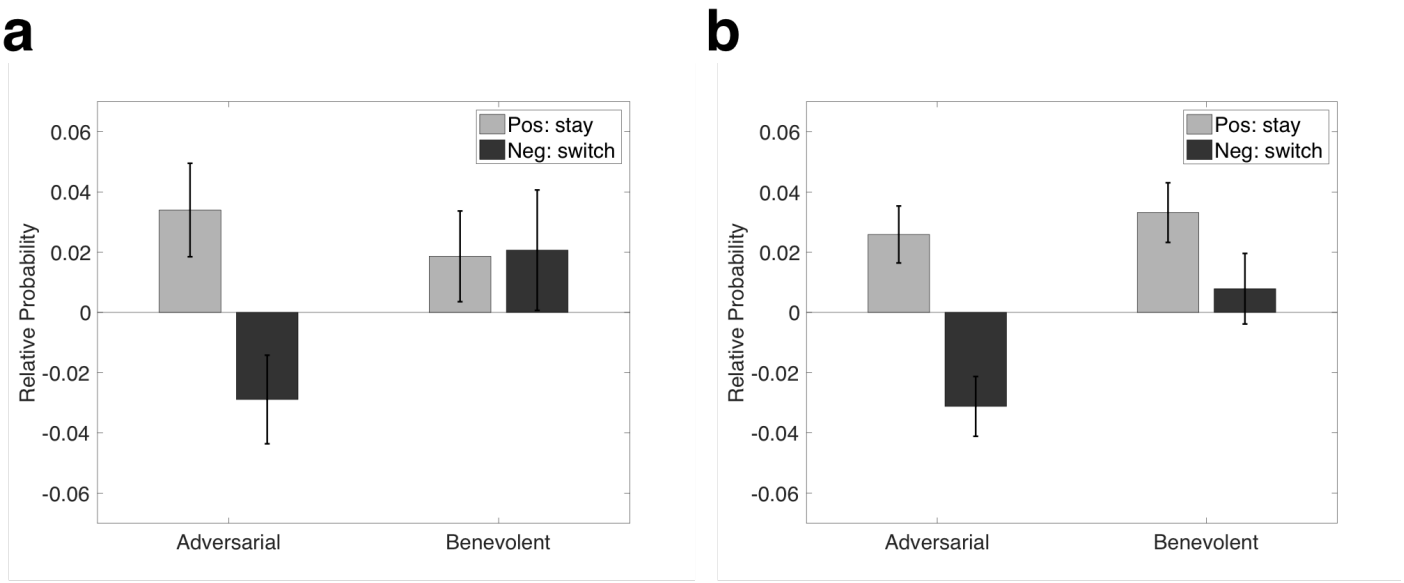


Figure S1.3.Win-stay Lose-shift. (a) Experiment 1 and (b) Experiment 2. Error bars represent across-subject standard error of the mean (SEM).

Task Performance

Performance data for choice behavior in the task is reported here for both Experiment 1 (N = 70) and Experiment 2 (N = 255).

Experiment 1. The mean proportion of trials where participants chose the more rewarding option was 0.794, with a standard deviation of 0.088.

Experiment 2. The mean proportion of trials where participants chose the more rewarding option was 0.785, with a standard deviation of 0.095.

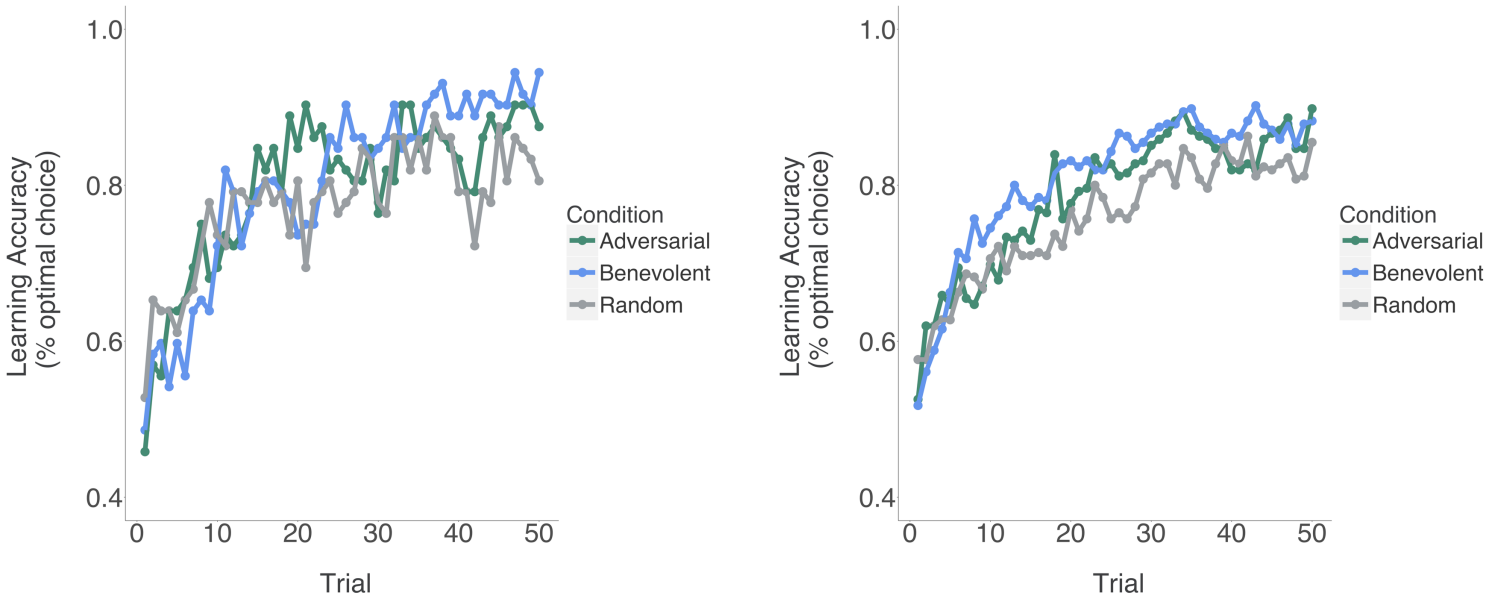


Figure S1.4. Percent optimal choice by trial and condition for Experiment 1 (left) and Experiment 2 (right).

Individual Differences Analyses

Initial Experiment

One of the broader questions motivating this research is how the environment shapes learning rate asymmetries. We addressed this question by performing an exploratory analysis where we investigated whether individuals' prior expectations influenced their beliefs about control. As proxies for prior expectations, we measured trait optimism and childhood socioeconomic status, under the assumption that these measures reflect ingrained beliefs about the prevalence of positive outcomes in the environment.

Method

Participants. 110 participants (49 female, 56 male, 5 unreported) from Amazon Mechanical Turk completed the two-alternative forced choice behavioral task outlined in the main text (Experiment 2) and self-report measures to assess trait optimism and socioeconomic status (SES). Participants completed some of the self-report measures during the same session as the behavioral task and the remainder of the measures during a separate session to avoid fatigue. A larger sample size was chosen compared to Experiment 1 in order to ensure that the sample would be sufficiently large after excluding for inaccuracy and incomplete responses. Ninety-five individuals completed some or all of the self-report measures but were excluded if they did not complete all of the self-report measures or did not meet the accuracy criterion in the behavioral task. Five additional participants were excluded from zip code analyses because they chose not to provide zip code information, entered an invalid code, or because median income data was not available for the location they entered. Self-report measure analyses for the Life Orientation Test – Revised (LOT-R) results included data from 89 participants and results from the zip code analyses included data for 84 participants. Participants gave informed consent, and the Harvard Committee on the Use of Human Subjects approved the experiment.

Self-Report Measures. Participants completed self-report measures in addition to the behavioral task, either directly after the task or in a separate online session. Participants completed the Life Orientation Test – Revised (LOT-R) (Scheier, Carver, & Bridges, 1994) to assess trait optimism/pessimism and the MacArthur Subjective Socioeconomic Status scale (Singh-Manoux, Marmot, & Adler, 2005). This questionnaire was revised to also collect objective measures of socioeconomic status including childhood zip code, current zip code, and current yearly income

(see Supplemental Materials). The sample completed additional measures that will not be reported here. Given our specific hypotheses about trait optimism and environmental markers of optimism and pessimism, we focus on LOT-R and SES results.

Results

In order to investigate whether optimistic and pessimistic biases are related to variability in beliefs about intervention, we tested the association between optimism scores and participants' mean belief in latent agent intervention (Fig. S1.5). We used a Spearman's rank-order correlation test due to the fact that LOT-R scores were optimistically skewed (*range*: 0-24; *mean*: 15.01; *median*: 17; *standard deviation*: 7.26). We found a significant correlation between belief in the latent agent across all conditions for negative feedback ($r_s = 0.256$, $p = 0.019$). Since benevolent agents cannot cause negative outcomes, we wanted to confirm that these results were indeed due to appropriate attributions in the adversarial and neutral conditions. We found a significant association between beliefs in the adversarial condition and optimism, ($r_s = 0.224$, $p = 0.034$), as well as for the neutral condition, ($r_s = 0.237$, $p = 0.025$), but not for the benevolent condition ($r_s = -0.021$, $p = 0.840$). A Fisher's *r*-to-*z* transformation revealed significant differences between both the adversarial and benevolent correlation coefficients ($z = 2.190$, $p = 0.029$) and the neutral and benevolent correlation coefficients ($z = 2.477$, $p = 0.013$). These results suggest that optimists are more likely to blame bad outcomes on external forces. It is interesting to note that we did not find any relationship between attribution of positive outcomes and trait optimism ($r_s = -0.045$, $p = 0.679$).

To explore a possible environmental source of asymmetries, we tested whether childhood socioeconomic status contributes to beliefs about agency. We used median income data from the most recent available data, the 2006-2010 American Community Survey (ACS) compiled by the University of Michigan Population Studies Center. We found that median income of participants' childhood zip code (range: $16,346 - $135,253 USD) was positively correlated with belief in the hidden agent for negative outcomes collapsed across the adversarial and neutral conditions ($r = 0.259$, $p = 0.019$) (Fig. S1.6). However, consistent with our findings on trait optimism, there was no significant relationship between agency beliefs about positive outcomes and childhood neighborhood income ($r_s = -0.089$, $p = 0.388$). A Fisher's r-to-z transformation confirmed that the correlations for negative and positive outcomes were significantly different ($z = 2.618$, $p = 0.009$).

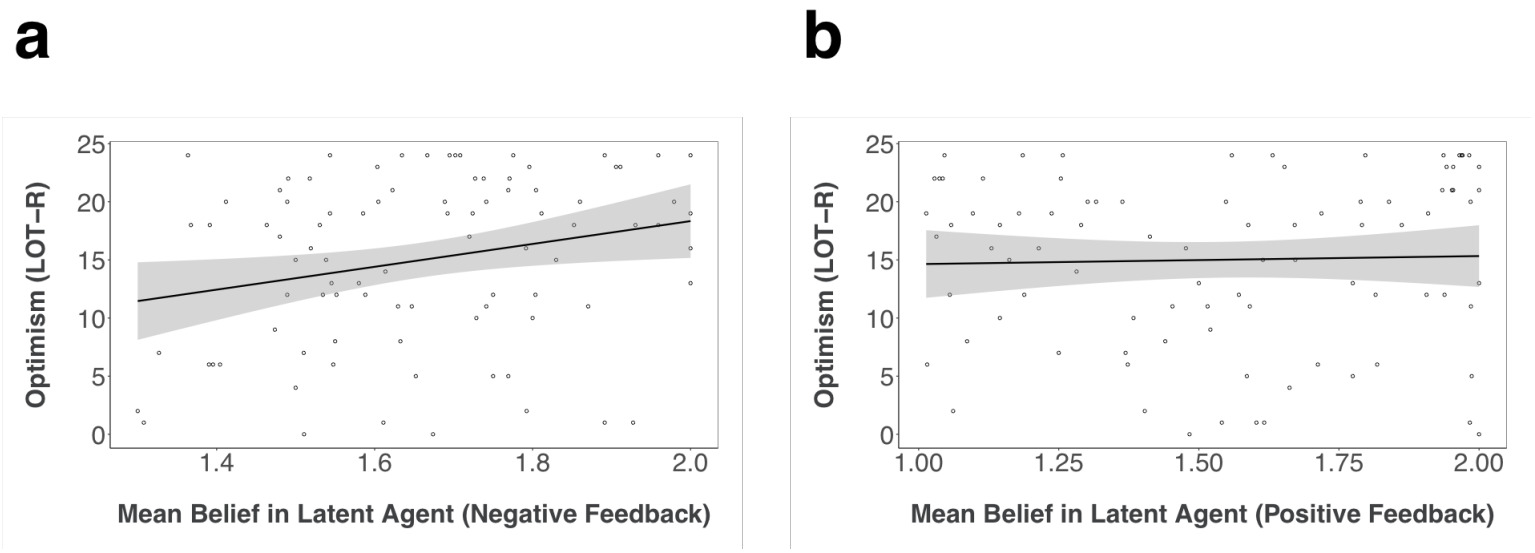


Figure S1.4. Optimism and beliefs about agency. LOT-R correlates with beliefs about latent agent intervention for (a) negative, but not (b) positive feedback. Negative feedback trials were included from adversarial and neutral conditions, and positive feedback trials were taken from the benevolent and neutral conditions.

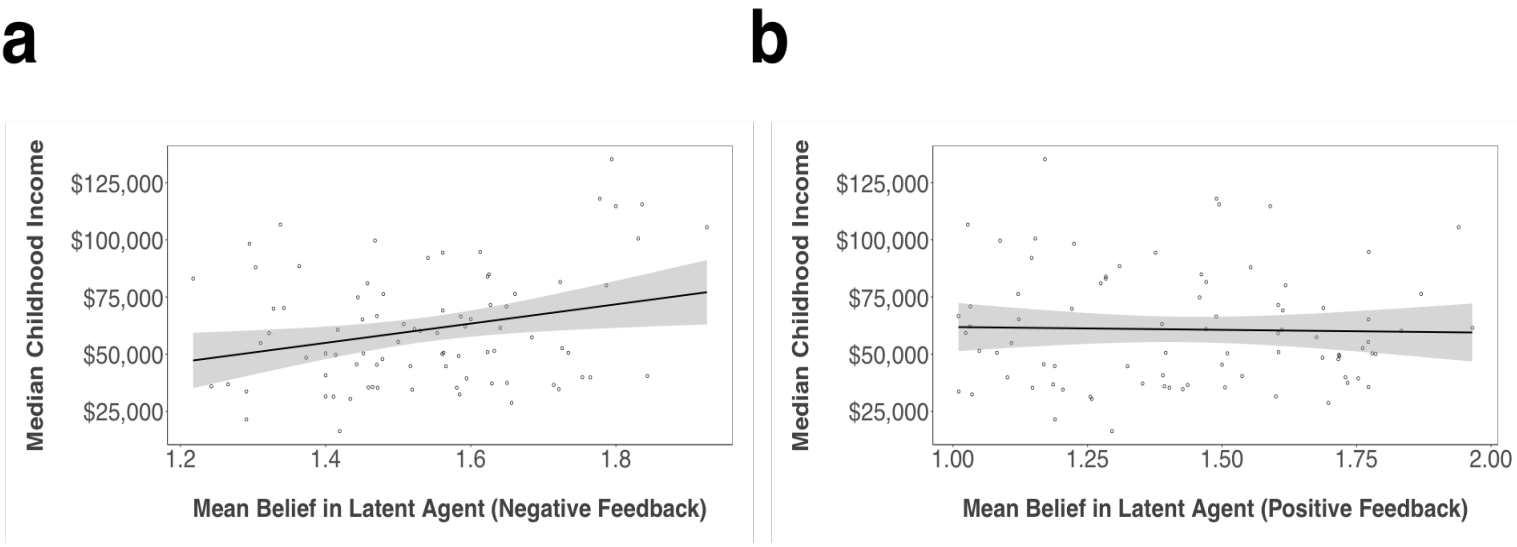


Figure S1.5. **Childhood environment and beliefs about agency**. Median income of the neighborhood where participants spent their childhood correlates with beliefs about latent agent intervention for (a) negative, but not (b) positive feedback. Negative feedback trials were combined across adversarial and neutral conditions, and positive feedback trials were pulled from benevolent and neutral conditions.

Pre-registered Replication

Due to small effect sizes and concerns about reproducibility, we attempted a pre-registered replication of the correlational individual differences analyses reported above. All procedures were identical to the original experiment. The replication was registered via the Open Science Framework: https://osf.io/3htpj/.

Method

Participants. 156 participants (74 female, 75 male, 8 other/unreported) from Amazon Mechanical Turk completed the two-alternative forced choice behavioral task outlined in the

main text (Experiment 2 and Initial Experiment above) and self-report measures to assess trait optimism and socioeconomic status (SES). The number of participants was determined using power analyses performed in RStudio (using package 'pwr'). In order to obtain 90% power, we determined that we would need 155 participants for the LOT-R correlation and 151 subjects for the SES correlation. We collected data for a total of 196 participants on Amazon Mechanical Turk in order to include 156 usable participants. Exclusion criteria were identical to all other experiments reported here: participants were excluded if they did not get a comprehension question correct, did not choose the higher-rewarded option for > 60% of all trials, did not have usable zip code data, or encountered technical difficulties submitting their full data set.

Self-Report Measures. Participants completed self-report measures in addition to the behavioral task, directly after the task. Participants completed the Life Orientation Test – Revised (LOT-R) (Scheier et al., 1994) to assess trait optimism/pessimism and the MacArthur Subjective Socioeconomic Status scale (Singh-Manoux et al., 2005). This questionnaire was revised to also collect objective measures of socioeconomic status including childhood zip code, current zip code, and current yearly income (see Supplemental Materials). The sample completed additional measures that will not be reported here. Given our desire to replicate our findings about trait optimism and environmental markers of optimism and pessimism, we focus only on LOT-R and SES results.

Results

We were unable to fully replicate all of our results from the initial experiment reported above. While some of our results did replicate (see below), we do not feel confident enough to draw any strong conclusions. Instead, we report the results here without further comment.

1. We find a replication of the correlation between belief in the latent agent across all conditions for negative feedback and LOT-R score ($r_s = 0.163$, $p = 0.043$).
2. We find no replication of the correlation between beliefs in the adversarial condition and LOT-R score ($r_s = 0.138$, $p = 0.085$).
3. We find no replication of the correlation between beliefs in the latent agent in the neutral condition and LOT-R score ($r_s = 0.069$, $p = 0.392$)
4. We find no replication of the non-correlation between beliefs in the latent agent in the benevolent condition and LOT-R score ($r_s = 0.196$, $p = 0.014$).
5. We find no replication of the non-correlation between attribution of positive outcomes and LOT-R score ($r_s = 0.206$, $p = 0.001$).
6. We find no replication of the significant correlation between median income of participants' childhood zip code and belief in the hidden agent for negative outcomes collapsed across the adversarial and neutral conditions ($r_s = 0.104$, $p = 0.198$)
7. We find a replication of the non-correlation between agency beliefs about positive outcomes and childhood neighborhood income ($r_s = 0.068$, $p = 0.398$).

References

Gershman, S. J. (2016). Empirical priors for reinforcement learning models. *Journal of Mathematical Psychology*, *71*, 1–6. http://doi.org/10.1016/j.jmp.2016.01.006

Gershman, S. J., Pesaran, B., & Daw, N. D. (2009). Human reinforcement learning subdivides structured action spaces by learning effector-specific values. *The Journal of Neuroscience : the Official Journal of the Society for Neuroscience*, *29*(43), 13524–13531. http://doi.org/10.1523/JNEUROSCI.2469-09.2009

Scheier, M. F., Carver, C. S., & Bridges, M. W. (1994). Distinguishing optimism from neuroticism (and trait anxiety, self-mastery, and self-esteem): A reevaluation of the Life Orientation Test. *Journal of Personality and Social Psychology*, *67*(6), 1063–1078. http://doi.org/10.1037//0022-3514.67.6.1063

Singh-Manoux, A., Marmot, M. G., & Adler, N. E. (2005). Does Subjective Social Status Predict Health and Change in Health Status Better Than Objective Status? *Psychosomatic Medicine*, *67*(6), 855–861. http://doi.org/10.1097/01.psy.0000188434.52941.a0

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

Supplementary Materials: Paper 2

Computational Models

Model Fitting. We used a softmax function to model choice probabilities, including a response stochasticity (inverse temperature) parameter and a "stickiness" parameter to capture choice autocorrelation (Gershman, Pesaran, & Daw, 2009). The Bayesian models were fit using maximum *a posteriori* estimation with empirical priors based on previous research (Gershman, 2016). Specifically, the prior distribution for the inverse temperature was $\beta \sim \text{Gamma}(4.82, 0.88)$, and for the stickiness parameter was $\rho \sim \mathcal{N}(0.15, 1.42)$.

Model Comparison. We used random-effects Bayesian model selection (Stephan, Penny, Daunizeau, Moran, & Friston, 2009) to compare models. This procedure treats each participant as a random draw from a population-level distribution over models, which it estimates from the sample of model evidence values for each model. We used the Laplace approximation of the log marginal likelihood to obtain the model evidence values. For our model comparison metric, we report the "protected exceedance probability" (*PXP*), the probability that a particular model is more frequent in the population than all other models under consideration. This is differentiated from an "exceedance probability" in that it considers the possibility that some differences in model evidence are due to chance.

Descriptive Model. In order to characterize learning rate asymmetries without committing to the assumptions of the Bayesian model, we fit a "descriptive" reinforcement learning model which updates reward probability estimates according to $\theta_{t+1} = \theta_t + \alpha_t(r_t - \theta_t)$, with separate learning rates for each combination of positive/negative outcome and the three experimental

conditions (benevolent, adversarial, random). We refer to this model in the main text as the *six-learning rate model*.

Bayesian Reinforcement Learning Model. This model considers a problem in which a decision-maker must choose an option $c \in \{1, \ldots, C\}$ based on experienced rewards. Each option is associated with an intrinsic reward distribution, but a latent cause can intervene (denoted $z = 1$) to generate a different reward distribution. With probability $P(z = 0)$, the decision-maker receives a reward from the intrinsic distribution, $P(r|c, z = 0)$, or with probability $P(z = 1)$, she receives a reward determined by the latent cause, $P(r|z = 1)$. The marginal distribution over reward given the decision maker's choice is thus

(1)

$$P(r|c) = P(r|c, z = 0)P(z = 0) + P(r|z = 1)P(r|z = 1).$$

In our setting, rewards are binary and distributed according to a Bernoulli distribution for each option, $Bern(\theta^c)$. Thus, absent intervention, a reward of 1 is drawn with probability θ^c, and a reward of 0 is drawn with probability $1 - \theta^c$.

For our experimental paradigm, we define the latent cause as another agent (the "latent agent") that can allot outcomes for the decision maker. We define 3 different latent agent types:

- Benevolent latent agent: produces a reward regardless of the decision maker's choice. Formally, $P(r = 1|z = 1) = 1$.

• Adversarial latent agent: produces no reward regardless of the decision maker's choice. Formally, $P(r = 0|z = 1) = 1$.

The decision maker does not know the true reward probabilities of her options. She has a prior belief that the unknown parameters θ^c are independently distributed according to $Beta(a, b)$, which are then updated from experience. The Beta distribution parameters can be fit to choice data though for simplicity we assume both are equal to 1 in the estimation procedure, corresponding to a uniform distribution.

After choosing an action c_t and observing reward r_t on trial t, the decision maker updates her estimate of the reward probability θ^c according to a reinforcement learning equation that incorporates inference over latent causes. With a Beta prior and a Bernoulli reward distribution, the Bayesian update rule takes the form $\theta_{t+1} = \theta_t + \alpha_t(r_t - \theta_t)$, where α_t is a parameter representing the learning rate that scales the reward prediction error. This learning rate is based rationally on beliefs about the outcome's two possible sources: the action's intrinsic reward and the latent agent's intervention. These beliefs jointly determine the extent to which the participant attributes feedback to each source. The learning rate is given by

(2)

$$\alpha_t = \frac{P(z_t = 0|r_t, c_t)}{N_t^c + a + b}$$

where $N_t^c \approx \sum_{\tau \in \{1,\ldots,t|c_\tau = c\}} P(z_\tau = 0|r_\tau, c_\tau)$ is the sum of past beliefs about latent agent non-intervention on trials when the same option was chosen. The denominator reflects the magnitude

of evidence about the intrinsic reward probability accumulated up to trial t (it is approximate because we assume for tractability that evidence provided by past feedback is not revised according to later information).

The learning rate's numerator is central to our present analysis: it encodes the degree to which feedback should be attributed to the intrinsic reward distribution rather than to the latent agent, and modulates the degree of learning based on whether feedback was positive or negative. The value of $P(z = 0|r, c)$ is stipulated by Bayes' rule:

(3)

$$P(z|r, c) = \frac{P(r|z, c)P(z)}{\sum_{z'} P(r|z', c)P(z')}.$$

This yields the following expressions which vary based on the combination of feedback and agent type:

Benevolent agent (negative feedback):

$$P(z = 0|r = 0, c) = 1$$

Benevolent agent (positive feedback):

$$P(z = 0|r = 1, c) = \frac{\theta^c P(z = 0)}{\theta^c P(z = 0) + P(z = 1)}$$

Adversarial agent (negative feedback):

$$P(z = 0|r = 0, c) = \frac{(1 - \theta^c)P(z = 0)}{(1 - \theta^c)P(z = 0) + P(z = 1)}$$

Adversarial agent (positive feedback):

$$P(z = 0|r = 1, c) = 1$$

where the probability of intervention $P(z = 1)$ is known. Subtly, the decision maker's inference about the latent agent's intervention depends on her existing estimate of θ^c.

The learning rate exhibits asymmetries depending on whether the latent agent tends to produce positive or negative outcomes. For example, when the agent is adversarial, positive outcomes can only come from the action itself, whereas negative outcomes are partly attributable to the external agent. Consequently, negative outcomes are less informative about the action's reward probability, corresponding to a lower learning rate.

We explored two possible models for this scenario: (1) the "fixed Bayesian" model, which treats the intervention probability as a free parameter, and (2) the "empirical Bayesian" model, which derives the intervention probability by averaging the participants' binary intervention judgments.

Fixed Bayesian Reinforcement Learning Model. Here, we fix the estimate of the latent agent's intervention probability, $P(z_t = 1)$, at 30% to mimic the instructions that participants receive. This is integrated into the Bayesian Reinforcement Learning Model described above.

Empirical Bayesian Reinforcement Learning Model. This model calculates the decision-maker's average intervention judgment and utilizes this value for the estimate of the intervention probability $P(z_t = 1)$, providing an individualized estimate of intervention for each participant. This is integrated into the Bayesian Reinforcement Learning Model described above.

Behavioral Analyses

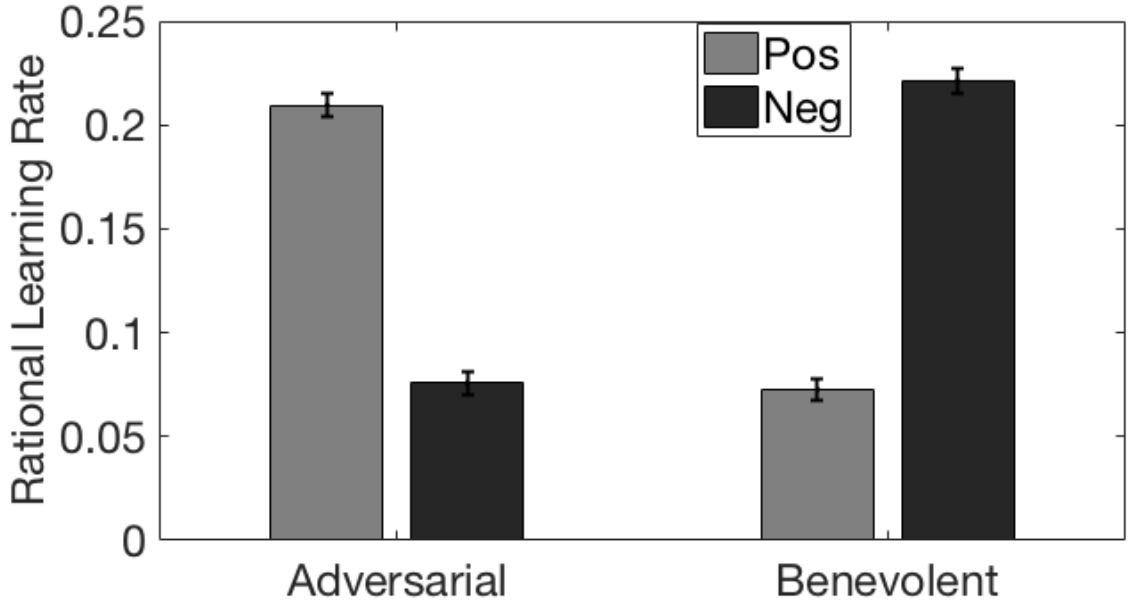


Figure S1. Learning Rates. The learning rates derived from the empirical Bayesian model and fit to participant data show distinct asymmetric learning of positive and negative outcomes by condition.

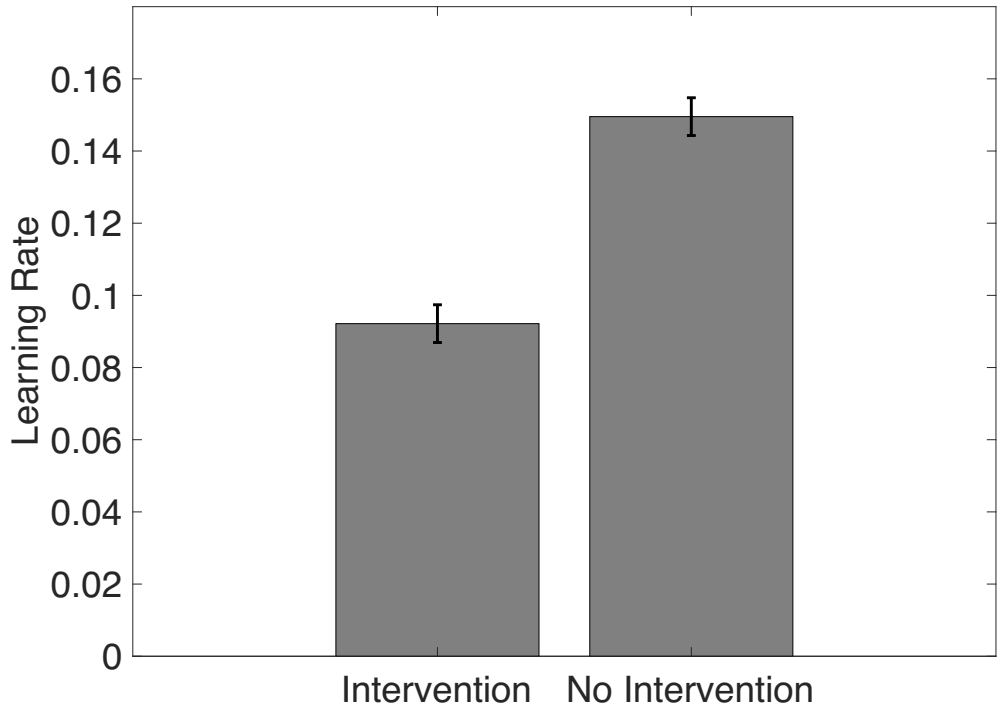


Figure S2.2. Learning & Hidden Agent Intervention. Learning rates are significantly higher for trials where the participants believe they caused an outcome ("No Intervention") compared to trials where participants believe the hidden agent caused an outcome ("Intervention").

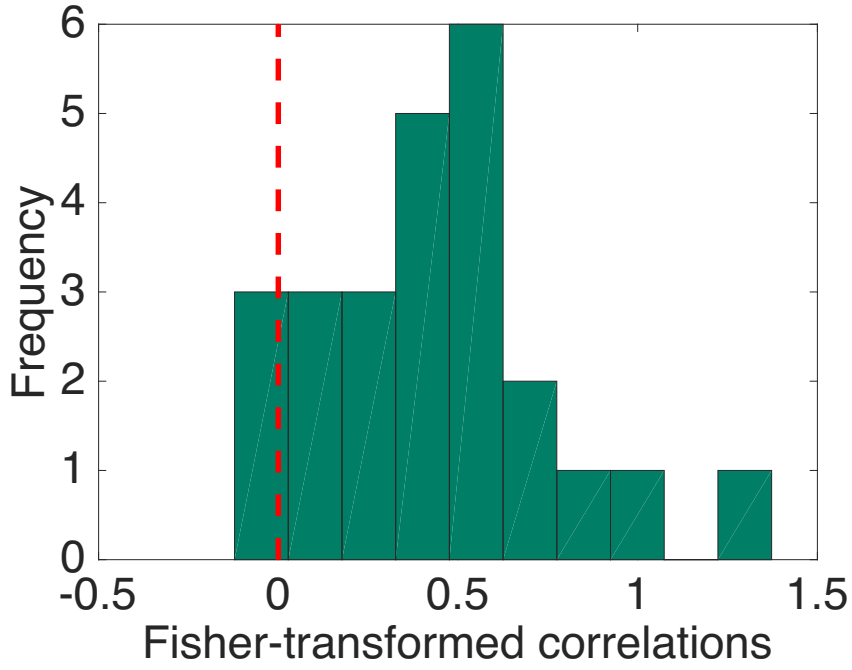


Figure S2.3. Model Estimates of Hidden Agent Intervention. Participants' judgments about intervention also show a significant median point-biserial correlation with the intervention predicted by the model, $r_{pb} = 0.424$, $p < 0.0001$.

References

Gershman, S. J. (2016). Empirical priors for reinforcement learning models. *Journal of Mathematical Psychology*, *71*, 1–6. http://doi.org/10.1016/j.jmp.2016.01.006

Gershman, S. J., Pesaran, B., & Daw, N. D. (2009). Human reinforcement learning subdivides structured action spaces by learning effector-specific values. *The Journal of Neuroscience : the Official Journal of the Society for Neuroscience*, *29*(43), 13524–13531. http://doi.org/10.1523/JNEUROSCI.2469-09.2009

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

Supplementary Materials: Paper 3

Computational Models

Model Comparison. We used random-effects Bayesian model selection (Stephan, Penny, Daunizeau, Moran, & Friston, 2009) to compare models. This procedure treats each participant as a random draw from a population-level distribution over models, which it estimates from the sample of model evidence values for each model. We used the Laplace approximation of the log marginal likelihood to obtain the model evidence values. For our model comparison metric, we report the "protected exceedance probability" (*PXP*), the probability that a particular model is more frequent in the population than all other models under consideration. This is differentiated from an "exceedance probability" in that it considers the possibility that some differences in model evidence are due to chance.

Adaptive Model. This model updates the weighting parameter from trial-to-trial based on the relative predictive accuracy between the two controllers. On each trial, the agent takes action(a) in response to stimulus (s) and receives reward (r). The agent considers two distinct generative models for the reward-generating process. The uncontrollable model assumes that rewards are independent of actions, whereas the controllable model assumes that rewards depend on actions. When rewards are binary, the models take the form of Bernoulli distributions governed by a latent parameter θ_{sa} in the controllable case and θ_s in the uncontrollable case. Because the Bernoulli parameter is unknown to the agent, it must be inferred. We will present inference for

the uncontrollable model; inference for the controllable model is identical except that the parameters are indexed by stimuli and actions.

According to 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:

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

Fixed Model. This model shares the same underlying information processing architecture as the adaptive model, but instead fits the weighting parameter as a free parameter.

Model Fitting. We used a softmax function to model choice probabilities, including a response stochasticity (inverse temperature) parameter. 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. 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*.

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.