Advantages of Confirmation Bias in Bayesian Inference

Atishay Narayanan [1] (atishay@princeton.edu)
[1]Department of Mathematics
Princeton University

Abstract

Learners are prone to a variety of biases. But is that necessarily a bad thing? Are there instances where biases could actually benefit a learner? In this paper, we examine confirmation bias, examined through the lens of an observer who is more prone to learn when exposed to confirming evidence. We use Python to model Bayesian Inference Agents with varying levels of bias and train them to learn both stationary and non-stationary Bernoulli distributions. In doing so, we examine whether having a degree of confirmation bias can help an Agent learn a distribution with less error.

Keywords: Bayesian Inference, Confirmation Bias

Introduction

Human beings, like all learners, exhibit bias when making observations. One such bias is confirmation bias, which oftentimes takes up different meanings depending on the context. Oftentimes, confirmation bias is used to describe situations where "information is searched for, interpreted, and remembered in such a way that it systematically impedes the possibility that the hypothesis could be rejected" (Oswald & Grosjean, 2004). However, the the complete refusal of alternative hypotheses is too strong of a definition for the scope of the paper, as in many situations there are several or possibly infinite hypotheses of varying likelihoods. In these situations, we look at the extension of the rule-discovery task to probabilistic environments explored in Klayman et. al. (Klayman & Ha, 1987). Furthermore, we focus on the idea of having a separate *positive factual learning* as discussed in Palimenteri et. al. (Palminteri, Lefebvre, Kilford, & Blakemore, 2017). The confirmation bias discussed from here on out follows from a reinforcement learning context, where we define confirmation bias as exhibiting a higher learning rate when presented with new observations that match the learner's hypothesis.

While many studies are focused on how to eliminate cognitive bias in studies (Pat, 2017), doing so operates under the assumption that bias poses a threat to learning. But bias is not necessarily harmful or even neutral in all cases. There are instances in which a degree of bias can actually aid a learner, as explored by Neuman et. al. (Neuman, Rafferty, & Griffiths, 2014). In this paper, Neuman et. al. shows that being biased towards positive results, i.e. having "Wishful Thinking", can lead to a higher total reward in some tasks. However, this study focuses on reward-based learning. We intend to focus on a different cognitive bias in an inference-based environment.

Inspired by this utilitarian approach towards bias, we aim to study whether confirmation bias can have benefits in learning the distribution of a Bernoulli random variable. We use Python simulations to model an Agent that observes a series of outcomes of a Bernoulli random variable and attempts to understand the underlying distribution of the variable via Bayesian Inference . We use the example of an unfair coin seen in (Griffiths, Tenenbaum, & Kemp, 2012) to help make the intuition clear. We add a degree of confirmation bias to our model allowing our agent to guess the outcome of a flip before it occurs, and over represent flips that the agent guesses correctly when it forms its Bayesian estimate of coin's true probability of landing on heads. Beyond the stationary Bernoulli distribution example that is a weighted coin, we also explore instances of Bayesian agents attempting to learn a non-stationary distribution.

We hypothesize that agents that exhibit confirmation bias will be better equipped to capture the Bernoulli distribution via Bayesian inference, especially in the non-stationary cases. Intuitively, this is because we reward observations with correct hypotheses, and as these become more likely as the estimation becomes more accurate, a biased agent should be capable of converging to the target distribution faster than an unbiased one.

The remainder of this paper is organized as follows: we begin by citing examples in literature that have already shown benefits to certain kinds of confirmation bias, then explore the mathematical background behind our Bayesian agents and how we simulate their observations. Then, we move onto our specific methods of implementing our model. Finally, we provide the results of our simulations and end with a discussion of our key findings.

Background

Confirmation Bias

Confirmation bias is often met with negative connotations. It is often associated with leading to errors, due to learners disregarding information that does not agree with their beliefs (Kappes, Harvey, Lohrenz, Montague, & Sharot, 2020). But as we see from Neuman et. al.'s work with cognitive bias towards positive results, learners who exhibit bias do not necessarily have worse results than unbiased learners. This is especially the case in the short-term time horizon (Neuman et al., 2014). In a similar study, results showed the same could be true for confirmation bias.

With respect to whether or not a belief was confirmed, Paliminteri et. al. found that "the positive factual learning rate... was significantly higher than the negative one" (Palminteri et al., 2017). Although we are applying it to a Bayesian inference environment rather than a reinforcement learning one, we aim to show that this still holds. To simulate this discrepancy, we introduce a scaling factor to our model (with respect to Paliminteri, α_c^+), which we refer to as the confirmation multiplier. When our Bayesian agent correctly guesses the outcome of a Bernoulli random variable and updates it's posterior, we will treat this as if the the Agent has witnessed the outcome multiple times, this multiple being controlled by the confirmation multiplier. This simulates a higher learning rate associated with positive outcomes.

Bayesian Inference

We utilize the Bayesian Inference framework used by Griffiths et. al. to estimate the probability of a hypothesis given prior and observed data in a continuous hypothesis space (Griffiths et al., 2012). The continuous hypothesis space here is given by $[0,1]$ and is the range of possible values of the Bernoulli random variable that the agent is attempting to estimate. For some set of data x, the probability that the hypothesis θ is true is given by equation (1).

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{\int p(x|\theta')p(\theta')d\theta'} \tag{1}$$

Specifically with respect to the situation of estimating the likelihood of a fair coin, we know from Griffiths et. al. that we end up with equation (2).

$$p(\theta|d) = \frac{(n_H+n_T+v_H+v_T+1)!}{(n_T+v_H)!(n_T+v_T)!}\theta^{n_H+v_H}(1-\theta)^{n_T+v_T} \tag{2}$$

Where v_H, v_T represent heads and tails observed a priori and n_h, n_t represent observed heads and tails during the trial. We have no interest in computing the above thousands of times across thousands of trials, and choosing the most likely value of θ as our posterior. So we turn to a reliable estimate that is tractable.

To choose a posterior value that represents the continuous distribution of hypotheses and the probability they occur, we turn to Maximum A Posteriori estimation (MAP) (Bassett & Deride, 2019) given in equation (3).

$$\theta_{MAP} = \frac{n_H+v_H}{n_H+n_T+v_H+v_T} \tag{3}$$

This gives us an unbiased agent's "best estimate" as to what the true probability is for a Bernoulli random variable such as a coin. To model what an agent with a degree of observation bias may estimate the true probability is, let the superscripts $+$ and $-$ denote whether an observed flip was guessed correctly or incorrectly prior to the observation, and let α_c^+ be our confirmation multiplier. Then equation (4) gives us our biased best estimate of θ.

$$\theta_{Bias} = \frac{\alpha_c^+ n_H^+ + n_H^- + v_H}{\alpha_c^+ n_H^+ + n_H^- + \alpha_c^+ n_T^+ + n_T^- + v_H + v_T} \tag{4}$$

In this way, we over represent cases where the agent correctly guesses an outcome when making an estimation of the true probability.

Non-Stationary Bernoulli Random Variables

In the real world, many distributions are not stationary; the true probability that an event is occurs can change over time depending on many factors. To account for this, we introduce a small drift factor sampled from a normal distribution $\mathcal{N}(0,\sigma^2)$ that we add to our ground truth after each observed event. If $X \sim \mathcal{N}(0,\sigma^2)$, then the PDF of X is given by equation (5).

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{x^2}{2\sigma^2}} \tag{5}$$

We reference our drift factor exclusively by its standard deviation σ. By adding this small value after each observation, we create a random walk that provides us with a non-stationary ground-truth distribution for our Bayesian agents to follow. We initialize our Bernoulli random variables uniformly at random for each trial and run each trial many times so our agents experience many varieties of these random walks.

Methods

All simulations were conducted using Python 3.11.14 on an M3 Macbook Pro running MacOS Sequoia version 15.6.1

The Bayesian Agent

Using Python and the Numpy library, we create an agent that has a prior Because we are using an MAP estimation to choose best possible posterior, the only data the agent needs to store is the count of observed flips and its prior. For the sake of simplicity, our experiments involve models exclusively initialized with a uniform prior, as if the prior observed was 1 head and 1 ail being flipped by a coin.

When our agent observes new information, we sample from its internal distribution to form a hypothesis. Then, we the coin is flipped, the agent observes the value and compares it to its hypothesis. If the hypothesis is correct, then the added the value $\alpha_c^+ \geq 1$ to the count of observations of that corresponding value. If the hypothesis is incorrect, then 1 is added.

Confirmation Multiplier

We decided on specific positive factual learning rate α_c^+ values via a parameter sweep, observing individual trials and accounting for how different levels of bias affect the models ability to converge to the ground truth distribution. We found that past $\alpha_c^+ = 2.0$, the agents became to volatile to accurately converge to a ground truth (see Figure 1). We did not employ any level of drift during the parameter sweep. Ultimately, we settled on $\alpha_c^+ = 1.00, 1.10, 1.25, 1.50, 1.75, 2.00$ for our experiments.

Drift

We decided on specific drift values via a parameter sweep, observing individual trials and taking note of how models with

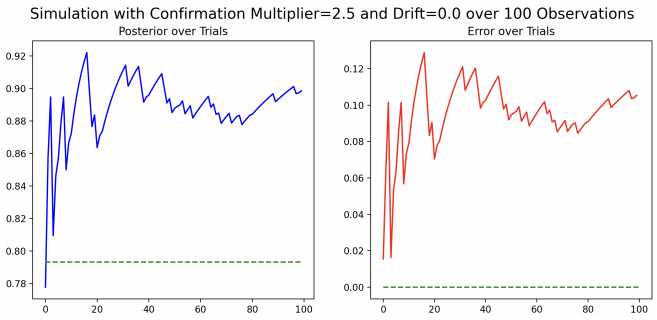

Figure 1: Example of trial involving agent with $\alpha_c^+ > 2.0$

different levels of confirmation bias responded to to difference values of σ. We noticed that in instances where we had $\sigma > 0.01$, the ground truth exhibited too much variance for any agent to follow due to instances of large divergences (see Figure 2). While this did not occur in all trials, the instances where it did skewed the MSE upwards.

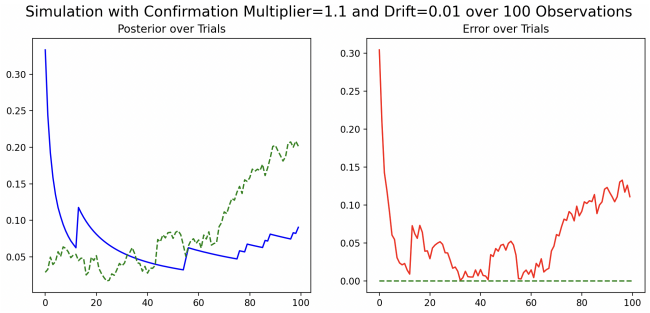

Figure 2: Example of non-stationary distribution with excessive drift

While still including an experiment with $\sigma = 0.01$ to provide an upper bound, we focused on more reasonable levels of drift such as $\sigma = 0.001$ and $\sigma = 0.005$. Using these, we were able to get non-stationary distributions that avoided excessive deviations in which all agents, regardless of bias, were unable to follow. Figure 3 helps visualize how this is more effective at providing a ground-truth for an agent to learn.

Simulations

We conducted two kinds of simulations: short term and long term. Short term simulations involve a series of 100 observations, and the error measured here gives us a better idea of how "fast" the agent is able to converge to the ground-truth distribution. The long term simulations include 1000 trials. These simulations are better interpreted as a measure of long term accuracy, testing the agents ability to both converge to and stick with the Bernoulli distribution, especially when it is non-stationary. Both simulations involve 1000 trials that each begin with a ground-truth Bernoulli distribution that is chosen uniformly at random from $[0,1]$.

After each observation, we sample a value from our drift distribution $\mathcal{N}(0,\sigma^2)$ and add it to our ground-truth. Each

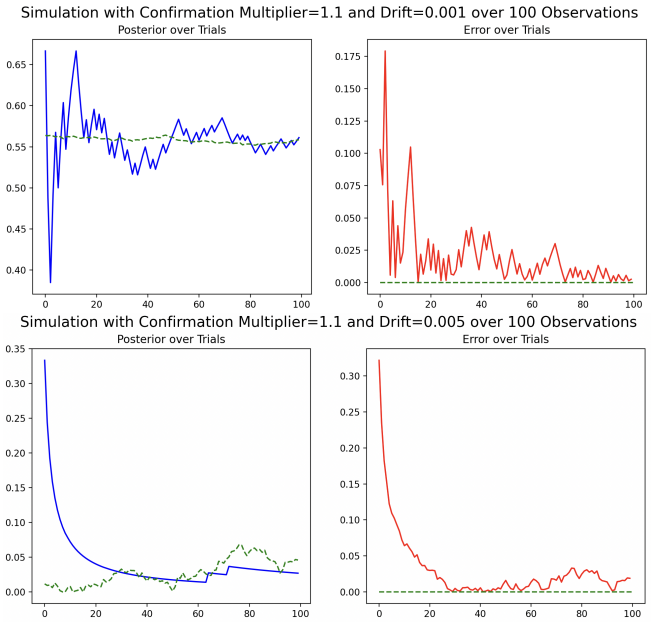

Figure 3: Examples of non-stationary distribution with reasonable levels of drift

time we add this drift, we clamp the value of the ground truth between 0 and 1 to ensure that we have a valid probability. Each observation uses the current probability for the event to occur.

Evaluation

To evaluate our Bayesian agents, we take the Mean Square Error (MSE) by taking the square of the difference of the agents posterior estimation and the true probability of the Bernoulli random variable at each observation and average it over each observation in a trial and then across all 1000 trials of an experiment. We calculate this with equation (5) and use sci-py's built in Standard Error Measure (SEM) to get the uncertainty of the measurement. For the sake of simplicity, we used Matplotlib's built-in bar chart function to make our graphs.

Results

The graphs for the short and long term simulations can be seen in Figures 4 and 5 respectively. The exact results for short term and long term experiments, organized by drift and then confirmation multiplier, can be seen in Tables 1 and 2 respectively.

Short Term

In the short term, agents with a small degree of confirmation bias manage to have lower MSE than the unbiased control agent. Specifically, agents with $\alpha_c^+ = 1.1$ exhibited lower MSE for experiments with drift $\sigma = 0.0$ and $\sigma = 0.01$, while performing within standard error during experiment with drift $\sigma = 0.001$ and $\sigma = 0.005$. Agents with with $\alpha_c^+ = 1.25$ performed particularly well in the experiment with low drift

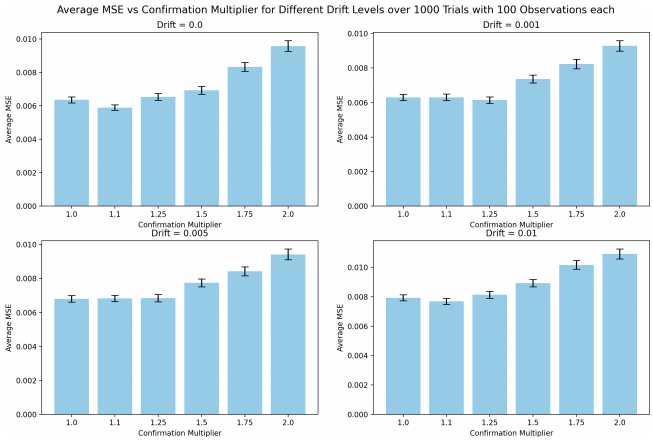

Figure 4: Short term distribution convergence experiment results

Table 1: Average MSE ($\pm$ SEM) over 1000 trials of 100 observations (short term).

Drift (σ)	1.00	1.10
0.000	$6.36 \times 10^{-3} \pm 1.86 \times 10^{-4}$	$5.89 \times 10^{-3} \pm 1.66 \times 10^{-4}$
0.001	$6.29 \times 10^{-3} \pm 1.81 \times 10^{-4}$	$6.30 \times 10^{-3} \pm 1.90 \times 10^{-4}$
0.005	$6.80 \times 10^{-3} \pm 1.90 \times 10^{-4}$	$6.83 \times 10^{-3} \pm 1.84 \times 10^{-4}$
0.010	$7.92 \times 10^{-3} \pm 2.07 \times 10^{-4}$	$7.68 \times 10^{-3} \pm 2.08 \times 10^{-4}$
Drift (σ)	1.25	1.50
0.000	$6.54 \times 10^{-3} \pm 1.86 \times 10^{-4}$	$6.93 \times 10^{-3} \pm 2.35 \times 10^{-4}$
0.001	$6.13 \times 10^{-3} \pm 1.86 \times 10^{-4}$	$7.37 \times 10^{-3} \pm 2.28 \times 10^{-4}$
0.005	$6.84 \times 10^{-3} \pm 2.15 \times 10^{-4}$	$7.74 \times 10^{-3} \pm 2.36 \times 10^{-4}$
0.010	$8.14 \times 10^{-3} \pm 2.38 \times 10^{-4}$	$8.92 \times 10^{-3} \pm 2.48 \times 10^{-4}$
Drift (σ)	1.75	2.00
0.000	$8.32 \times 10^{-3} \pm 2.65 \times 10^{-4}$	$9.58 \times 10^{-3} \pm 3.21 \times 10^{-4}$
0.001	$8.24 \times 10^{-3} \pm 2.76 \times 10^{-4}$	$9.29 \times 10^{-3} \pm 2.98 \times 10^{-4}$
0.005	$6.13 \times 10^{-3} \pm 1.86 \times 10^{-4}$	$6.30 \times 10^{-3} \pm 1.90 \times 10^{-4}$
0.010	$10.17 \times 10^{-3} \pm 2.96 \times 10^{-4}$	$10.91 \times 10^{-3} \pm 3.33 \times 10^{-4}$

($\sigma = 0.001$) and remained within the SEM for all other experiments.

However, all agents with any level of confirmation bias higher $\alpha_c^+ = 1.25$ underperformed compared to the unbiased control. While the gap shrank as drift increased, the highly biased agents never fell within the SEM of the unbiased agents during any experiment. Too much bias, it seems, does more harm than good regardless of the time-scale.

Long Term

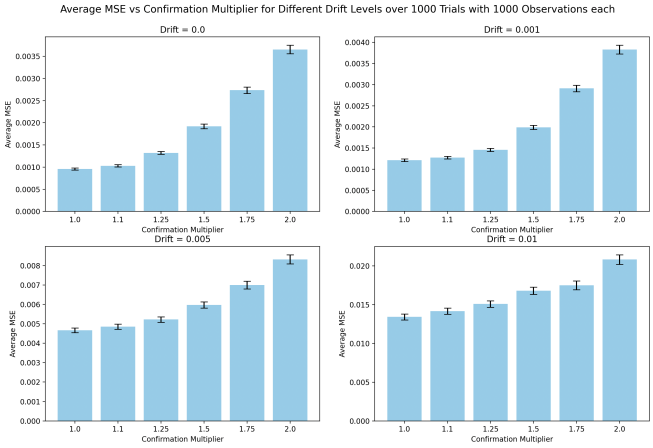

Figure 5: Long term distribution convergence experiment results

Table 2: Average MSE ($\pm$ SEM) over 1000 trials of 1000 observations (long term).

Drift (σ)	1.00	1.10
0.000	$0.96 \times 10^{-3} \pm 0.25 \times 10^{-4}$	$1.03 \times 10^{-3} \pm 0.26 \times 10^{-4}$
0.001	$1.21 \times 10^{-3} \pm 0.30 \times 10^{-4}$	$1.28 \times 10^{-3} \pm 0.32 \times 10^{-4}$
0.005	$4.67 \times 10^{-3} \pm 1.16 \times 10^{-4}$	$4.85 \times 10^{-3} \pm 1.31 \times 10^{-4}$
0.010	$13.42 \times 10^{-3} \pm 3.73 \times 10^{-4}$	$14.16 \times 10^{-3} \pm 3.98 \times 10^{-4}$
Drift (σ)	1.25	1.50
0.000	$1.32 \times 10^{-3} \pm 0.34 \times 10^{-4}$	$1.92 \times 10^{-3} \pm 0.52 \times 10^{-4}$
0.001	$1.46 \times 10^{-3} \pm 0.37 \times 10^{-4}$	$1.99 \times 10^{-3} \pm 0.51 \times 10^{-4}$
0.005	$5.22 \times 10^{-3} \pm 1.46 \times 10^{-4}$	$5.97 \times 10^{-3} \pm 1.62 \times 10^{-4}$
0.010	$15.09 \times 10^{-3} \pm 4.21 \times 10^{-4}$	$16.81 \times 10^{-3} \pm 4.74 \times 10^{-4}$
Drift (σ)	1.75	2.00
0.000	$2.74 \times 10^{-3} \pm 0.72 \times 10^{-4}$	$3.66 \times 10^{-3} \pm 0.94 \times 10^{-4}$
0.001	$2.91 \times 10^{-3} \pm 0.76 \times 10^{-4}$	$3.83 \times 10^{-3} \pm 1.02 \times 10^{-4}$
0.005	$6.99 \times 10^{-3} \pm 1.97 \times 10^{-4}$	$8.32 \times 10^{-3} \pm 2.38 \times 10^{-4}$
0.010	$17.50 \times 10^{-3} \pm 5.54 \times 10^{-4}$	$20.83 \times 10^{-3} \pm 6.22 \times 10^{-4}$

In the long term, no degree of bias shows any benefit. Across all 4 levels of drift, all 5 biased agents are outperformed by the the unbiased agent. However, it is important to note that the magnitude of the advantage diminishes as we add more drift. At $\sigma = 0.005$ and $\sigma = 0.01$, the unbiased agent falls within the SEM of the agents with $\alpha_c^+ = 1.1$. This suggest that the short term benefits of confirmation bias are mitigated in the long term. All agents with $\alpha_c^+ \geq 1.25$ performed poorly relative to the unbiased agents, especially in instances without or with low drift.

Discussion

Based on the results, our initial hypothesis is supported by the short term experiments. In the short term, having a small degree of confirmation proved to provide an advantage for agents. We speculate that this is because having the larger parameter α_c^+ in the MAP provides a greater factor to move the internal distribution, and because this factor is added in for confirming observations, once an agent has converged to the distribution, it can express it better. We notice that the the higher factors of confirmation bias underperformed, and this can be attributed to having too much volatility, making it difficult to stick with a distribution once the agent has confirmed. This is supported by the fact that in instances with higher drift, we observe a greater difference in MSE for highly-biased agents relative to unbiased and low-bias agents.

The data from the long term experiments do no support our hypothesis. The unbiased agent performed strictly better across all scenarios, leading us to believe that confirmation bias cannot be helpful in the long term, at least with respect to Bayesian inference. It seems that the advantages of converging to the target distribution faster than the unbiased agents do not outweigh the long term variance of the posterior estimation, even with the low-bias agents that proved to be at par or better than the unbiased agent in the short term. We noticed that the relative advantage of the unbiased agent seems to decrease as drift increases, as the MSE lands closer together in instances of higher drift. These results suggest that the advantages and disadvantages of biases can change based on the volatility of the environment. Overall, the data suggests that confirmation bias can provide some advantages, but these are time-scale dependent and serve as a detriment if the bias is too great.

Further Study

There are several ways this study can be enhanced or taken in a different direction. An important issue that was not accounted for is that as the number of observations in a trial increases, the relative impact of later observations on the agent's posterior estimation decreases, but we do not decrease the drift to account for this. This makes it difficult for agents, regardless of bias, to stay converged to a non-stationary Bernoulli distribution. Introducing a decay factor for σ could help with this, and choosing this factor should involve finding a precise value to keep the rate of change of the MAP estimation constant as a trial progresses.

Another area to explore would be the opposite of the effect studied in this paper. While Paliminteri et. al. suggests that there is a higher learning rate for confirming observations, intuitions suggests that this is not always the case. Learners often learn best from their mistakes. Replicating this experiment with higher learning rate for observations with incorrect hypotheses may provide insight as to whether that translates into the case of Bayesian inference.

Conclusion

In this paper, we examined whether confirmation bias can benefit a learner. Based on a model of Bayesian inference, we created agents that observed a series of outcomes for stationary and non-stationary Bernoulli random variables. These agents made predictions prior to observing the outcome, and agents with a degree of bias weighed examples that confirmed their hypothesis more when estimating the posterior distributions of these random variables.

We hypothesized that having a degree of confirmation bias would help and agent converge to the ground-truth distribution faster and stick with non-stationary distributions through their volatility. We saw some evidence of faster convergence speed in the short term for agents with low degrees of confirmation bias. However, the data was overwhelmingly in favor of the unbiased agents over the long term simulations. All biased agents exhibited much higher rates of error over the course of the long term simulation compared to their unbiased counterparts. This suggests that while having a small degree of confirmation bias can have nominal benefits in the short term, these advantages are mitigated in the long term. In the context of cognition as a whole, these results show that in certain cases, bias provides an advantage to learners.

Acknowledgments Thank you to Elizabeth Mieczkowski for guiding me through the Cognitive Science research process.

References

Bassett, R., & Deride, J. (2019). Maximum a posteriori estimators as a limit of bayes estimators. *Mathematical Programming*, *174*(1), 129–144.

Griffiths, T. L., Tenenbaum, J. B., & Kemp, C. (2012). Bayesian inference. *The Oxford handbook of thinking and reasoning*, 22–35.

Kappes, A., Harvey, A. H., Lohrenz, T., Montague, P. R., & Sharot, T. (2020). Confirmation bias in the utilization of others' opinion strength. *Nature neuroscience*, *23*(1), 130–137.

Klayman, J., & Ha, Y.-W. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological review*, *94*(2), 211.

Neuman, R., Rafferty, A., & Griffiths, T. (2014). A bounded rationality account of wishful thinking. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 36).

Oswald, M. E., & Grosjean, S. (2004). Confirmation bias. *Cognitive illusions: A handbook on fallacies and biases in thinking, judgement and memory*, *79*, 83.

Palminteri, S., Lefebvre, G., Kilford, E. J., & Blakemore, S.-J. (2017). Confirmation bias in human reinforcement learning: Evidence from counterfactual feedback processing. *PLoS computational biology*, *13*(8), e1005684.

Pat, C. (2017). Cognitive bias mitigation: becoming better diagnosticians. In *Diagnosis* (pp. 257–287). CRC Press.