Learning From Small Samples: A Machine Learning Model of Belief Polarization

by

Donghyuk Kim

A Thesis

Submitted to the Faculty

of the

WORCESTER POLYTECHNIC INSTITUTE

In partial fulfillment of the requirements for the

Degree of Master of Science

in

Computer Science

by

DEC 2022

APPROVED:

Professor Daniel Reichman, Major Thesis Advisor

Professor Craig Shue, Head of Department

Abstract

Why do people polarize? A common assumption is that people observe divisive information and form divisive beliefs. However, empirical studies indicate people can polarize even when they observe similar information. One possible reason is that people have cognitive biases and think irrationally. For instance, confirmation bias suggests prior beliefs influence how people interpret new evidence. We propose an alternative explanation for belief polarization. In an increasingly connected world, people are exposed to an abundance of information concerning a multitude of subjects. However, processing information is costly, so people may rely on small samples to form beliefs. Since small samples do not accurately represent the population and are variable, people may draw divergent images of the objective reality. First, we support our hypothesis using evidence from cognitive science literature. Then, we create a belief polarization model to test our hypothesis. We explore a unique approach and design a belief formation model based on machine learning. We propose new evaluation metrics for polarization and run simulations to observe how sample size affects polarization in various learning settings. Our results align with our hypothesis that, under basic assumptions, small sample reliance increases polarization. We offer practical suggestions for mitigating polarization based on our findings.

Acknowledgements

I would like to express my gratitude to my advisor, Dr. Daniel Reichman, who has provided invaluable advice and guidance during my study.

I am also deeply grateful to my family for their support, appreciation, and encouragement of my academic achievements. My journey would not have been possible without them.

Above all, I thank the Lord for my life, my health, my wisdom, and the strength you give me to face every day with hope.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Significance of Belief Polarization

In 1998, the *Lancet* published an article linking the MMR vaccine to autism. Even though the report has been heavily criticized and later retracted, it triggered an uptake in vaccine hesitancy around the world, an issue that lingers to this day [17]. In September 2020, for instance, Americans were split on the question of whether they would get vaccinated once the coronavirus vaccine is made available [37]. The controversy surrounding vaccines can lead to unfortunate outcomes as people may potentially miss out on life-saving treatments.

Belief polarization, such as people holding divergent views about vaccination, is a common empirical phenomenon in society. Are humans responsible for climate change? Do harsher punishments reduce crime? Will the new tax system improve the economy? People often disagree sharply on consequential issues that guide our lives. Polarization can be dangerous as when one group entirely refuses to consider another group's perspectives, it can thwart democratic solutions to social problems. Studies even suggest political polarization can increase the risk of violent conflicts within and between sovereign states [15].

The common wisdom among researchers is that American society is becoming increasingly polarized. Over the past several decades, a growing share of Americans say there are major differences in what the parties stand for, argue that partisan disagreements extend beyond policies to basic facts, and tend to vote along party lines [13]. Consequently, polarization has become a hot topic for social scientists.

1.2 Common Explanations for Polarization

What causes belief polarization? One possible factor is that people often obtain information from divisive sources. The human tendency to interact with like-minded others [31], as well as recommendation algorithms that cater to user preferences [4][38], can create echo chambers in which people positively reinforce and strengthen their shared beliefs to the extreme [18]. In addition, the rising amount of misinformation online [3] can widen the discrepancy between beliefs. In a setting where two people are exposed to contrasting information, polarization is a natural outcome.

However, divergent information does not fully explain why people polarize. There is evidence that people are reasonably good at detecting misinformation [2] and even echo chambers are not sealed off from unbiased information [34]. More importantly, empirical studies indicate that people can polarize even when they consider similar information [16]. Researchers often attribute this to cognitive biases that influence how people process new evidence. For example, confirmation bias [30] suggests people interpret information in a manner that validates their prior beliefs. However, this explanation is also incomplete as it does not address why people have contrasting priors in the first place.

1.3 Hypothesis

We propose an alternative explanation for belief polarization. In an increasingly connected world, people are exposed to an abundance of information concerning a multitude of subjects. However, processing information is costly, so people may rely on small samples to form beliefs. Since small samples do not accurately represent the population and are variable, people may draw divergent images of the objective reality. For instance, consider people who browse news articles online. Within a short period of time, they can be exposed to hundreds of issues from a wide range of topics. However, there is a limit to how much information people can process as reading and understanding articles take time and effort. Therefore, people may only read the headlines or just skim through the articles. Even when people read the articles, they are unlikely to research further and look for related articles. In other words, people tend to rely only on a few samples of information before forming opinions about a particular issue. Assuming people have to form numerous beliefs and that samples are costly, small sample reliance may be an optimal strategy for belief formation. For a particular subject, however, small samples may neglect different perspectives of the issue and lead people to form extreme beliefs. We hypothesize that reliance on smaller samples leads to greater polarization.

1.4 Thesis Outline

In chapter 2, we support our hypothesis using evidence from cognitive science literature. We discuss people's tendency to rely on small samples, the theories on why people rely on small samples, and how it translates to belief polarization. In chapter 3, we discuss computational models of belief polarization. We discuss prior models, formalize the definition of polarization, and design our model. We explore a novel approach to modeling belief polarization using machine learning. To the best of our knowledge, our work is the first to provide a machine-learning framework for empirical analysis of belief polarization. In chapter 4, we simulate agents to observe how sample size affects polarization in various learning settings. In chapter 5, we conclude the work with a summary of our findings and discussions on potential ways to mitigate polarization.

Chapter 2

Small Sample Reliance

2.1 Law of Small Numbers

Our hypothesis that people's reliance on small samples leads to polarization is grounded on Tversky and Kahneman's "law of small numbers" [40]. The law of small numbers suggests that people tend to overestimate the stability of estimates that come from small samples. That is, people erroneously believe small samples randomly drawn from the population are highly representative and are "similar to the population in all essential characteristics." If small samples can depict divergent views of reality, and people believe those views to be representative, their beliefs will diverge.

Consider the following example. Suppose the counties in which the cancer rate is lowest are mostly rural and sparsely populated counties in the Midwest. Based on this fact, one may logically deduce that this is so because people in rural areas generally have a low-stress lifestyle. Next, suppose that the counties in which the cancer rate is highest are also mostly rural and sparsely populated counties in the Midwest. Based on this fact, one may logically deduce that this is so because people in rural areas live in poverty. In this example, the counties in which the cancer rate is the highest and the lowest are both mostly rural areas because of the small population. Smaller populations are more variable as they are more sensitive to extreme samples. That is, people with unusually high rates or unusually low rates of cancer affect the statistic more. Therefore, people may form opposing ideas about cancer rates in rural areas depending on which statistics they observe. In other words, small samples can create polarizing beliefs.

2.2 Evidence of Small Sample Reliance

Empirical evidence indicates people tend to rely on small samples, and researchers have attributed various behavioral phenomena to that fact. One example is the underweighting of rare events. Several decision-making studies [22][23][35] have conducted experiments where participants choose between two gambles with unknown payout distributions. Although the participants were encouraged to sample until they felt confident about which gamble is better, most made less than 10 draws from each distribution before deciding which gamble to play. Researchers observed that because people relied on small samples, there were discrepancies between the objective probabilities of outcomes and the relative frequencies the participants experienced. For instance, many participants did not even encounter the rare event, so they behaved as if rare outcomes had less impact than their actual probabilities. On the other hand, people generally overestimated the probability of common events.

Another behavioral phenomenon implied by small sample reliance is the bias toward probability matching. Probability matching describes a decision strategy in which predictions of class membership are proportional to the class probability. For instance, one study [6] asked the participants to guess which color would appear on the screen next. If they made a correct guess, they would gain some amount of reward, and if they made an incorrect guess, they would lose the same amount. Suppose the color red appears with a probability of 0.7 and the color blue appears with a probability of 0.3. The optimal strategy would be to predict red in every single trial. However, participants generally exhibited probability matching. Researchers suggest this can be explained by people's reliance on small samples. For instance, if people only considered the past four examples and predicted the common color as the next color, they would predict red and blue with probabilities close to 0.7 and 0.3, respectively.

2.3 Why Small Samples?

Why do people rely on small samples? Some researchers argue people may rely on small samples due to their advantages. For instance, small samples can benefit the functioning of organisms, such as detecting changes in the environment [24], detecting correlations [26], foraging [5], and language learning [14]. In another school of thought, researchers argue that humans' cognitive limitations restrict people from considering too many samples. In 1956, for instance, the famous psychologist George Miller established the notion of a working memory capacity [32]. In his seminal work, Miller suggests that an average adult's short-term memory can process only around seven chunks of meaningful information at a time. Since then, a popular approach in research has been to understand human cognition as the optimal use of limited resources. The modeling paradigm is sometimes referred to as "resource-rational analysis" [29]. In this paradigm, the brain represents the world as a trade-off between accuracy and metabolic cost. In many models of cognition, researchers assume acquiring and processing samples are costly. Here, the cost may represent time, metabolic energy, memory, or even effort [43]. Large samples are necessary to accurately approximate the probability distribution. However, samples are costly. Therefore, people face a trade-off between accuracy and sample size.

One study [43] explicitly studies this trade-off and asks, "if people are making decisions based on samples – but as samples are costly – how many samples should people use to optimize their total expected or worst-case reward over a large number of decisions?" For instance, consider choosing a route to avoid traffic. How many possible arrangements of traffic across the city should one consider before deciding whether to turn left or right at the next intersection? Clearly, one should not pause at the intersection for hours to consider all the possibilities. Yet, one should also not make a random decision without any consideration. The researchers suggest that in a such setting where people face a trade-off, making many quick but locally sub-optimal decisions based on a few samples may be a globally optimal strategy over long periods. Similarly, we hypothesize that people form many beliefs based on a few samples, which may be locally sub-optimal. For one particular belief, however, people's beliefs may be inaccurate and polarize with others' beliefs.

Chapter 3

Model

3.1 Prior Models

The purpose of a computational model is to capture, even roughly, the plausible underlying mechanisms of empirical phenomena. Designing a computational model for polarization is a difficult task. First of all, the term polarization is not one unambiguous concept but a blurred cluster of concepts. In what sense do we mean by polarization? In the research literature, a range of very different social dynamics is lumped together under the term polarization. In addition, how humans learn and form beliefs is a complex phenomenon involving infinitely many variables.

3.1.1 Bayesian Model

One of the more popular approaches has been to use Bayesian inference to model human reasoning [42][39]. In Bayesian inference, an agent updates the probability of a hypothesis as it receives new evidence according to Bayes' theorem:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}. \tag{3.1}$$

The prior hypothesis $P(A)$ represents original belief, the likelihood $P(B|A)$ represents the probability distribution of the observed data given the original belief, and the posterior $P(A|B)$ represent the updated belief. Researchers following Bayesian inference typically define polarization as when two agents observe the same evidence but update their hypotheses in the opposite direction. For instance, consider a problem where two people observe the same data d that bear on some hypothesis h_1. Let P_1 and P_2 be probability distributions that capture two people's beliefs. Then, polarization occurs when

$$[P_1(h_1|d) - P_1(h_1)][P_2(h_1|d) - P_2(h_1)] < 0. \tag{3.2}$$

That is, polarization occurs when one person's belief in h_1 increases while the other person's belief in h_1 decreases, despite observing the same data. Following this formulation, studies explore various factors such as views on the reliability of the evidence sources [25], selective sharing of information [8], and bounded rationality [36].

Although Bayesian models have been the standard for studying human inference due to their flexibility, it has some disadvantages. Primarily, Bayesian inference requires researchers to translate subjective prior beliefs into mathematical prior. However, there is no standard method for doing so, and expressing priors to account for various factors has become a convoluted task. In addition, there is evidence that Bayesian updating is not always consistent with human behavior [11].

3.1.2 Learning Theory Model

A recent work titled *Belief Polarization in a Complex World* by Haghtalab, Jackson, and Procaccia [19] has taken a novel approach by using learning theory to model belief polarization. Specifically, the work follows the *probably approximately correct learning* framework [41], which is used for mathematical analysis of machine learning. In this framework, a learning agent observes train samples from some distribution $\mathcal{D}$ to select a generalization function h (hypothesis) that approximates the target function f. For a given train sample (x, y), x represents a set of information and y represents the belief. In this view, the target function $f : X \rightarrow Y$ is the objective reality that correctly maps the information to the belief. The agent selects the hypothesis h from a certain class of functions H known as the hypothesis set, which represents a space of possible hypotheses for mapping inputs to outputs. An expressive hypothesis set allows model agents to learn complex functions that capture complex relationships. The agent selects $h \in H$ that minimizes some generalization error $err_{\mathcal{D}}(h)$, such as the squared error $(h(x) - f(x))^2$.

To define polarization, the researchers construct a binary classification task with the labels $Y = \{-1, 1\}$. They define the *disagreement* between the two learned hypotheses as $Pr_{x \sim \mathcal{D}}[h(x) \neq h'(x)]$. Then, they define polarization as when the disagreement between two agents is disproportionately large compared to the difference between the distributions they acquired the samples from.

One of the models that the researchers propose in this study is the *complex objective model*, which argues that cognitive bounds restricting humans' model complexity can increase polarization. For example, a certain problem may require 10 dimensions of factors to understand, but it may be costly for a person to consider more than 7 of those dimensions. When different agents consider a different subset of 7 dimensions, their belief functions may end up drastically different. The complex objective model requires that agents not only minimize the generalization error $err_{\mathcal{D}}(h)$, but also the model complexity cost $\phi(h)$. The paper utilizes theoretical arguments based on learning theory to suggest it is possible to construct a task of minimizing both $err_{\mathcal{D}}(h)$ and $\phi(h)$ such that even when two agents observe train sets from the same distribution, the disagreement between the two optimally learned hypotheses will be high.

Our work aims to expand on Haghtalab, Jackson, and Procaccia's method of using machine learning to model polarization. Similar to their work, we study how cognitive limitations can increase polarization. However, rather than designing a theoretical model of polarization, we take an empirical approach and explicitly instantiate a machine-learning framework that emulates how humans learn.

3.2 Model Specification

Designing a belief polarization model involves specifying three main components: the learning agent, the task, and the measure of polarization. We rely on evidence from cognitive science to create a model that emulates how humans form beliefs.

3.2.1 Task and Learning Agent

The world that people experience is dominated by approximately linear relationships [21]. Meaning, linear models provide a nearly complete summary of the environment in many domains. In addition, linear models capture the essential characteristics of how humans make decisions [1]. For many decision tasks, people simply integrate information on numerical scales, weigh them, and add them up. Research has shown linear models often capture most of our behaviors. Even expert judges forming beliefs, such as a psychologist making a diagnosis, resemble a linear model [21]. Consequently, decision-making studies often assume a linear view of human cognition [27][10] [12].

Following the assumption that many relationships in the world are linear, we represent our machine learning task as a multiple linear regression problem with the form:

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_n x_n + \epsilon. \tag{3.3}$$

In equation 3.3, the explanatory variables x_i represent factors that contribute to a belief, the coefficients β_i represent the importance weight of each factor, and the noise ϵ represents the irreducible noise in nature. The response variable y represents the belief. We represent the belief as a real number value. This approach aligns with the practice of representing beliefs as being distributed on a normalized spectrum along one dimension [9]. For our model, we generate the target distribution by selecting the explanatory variables, the coefficients, and the irreducible noise from a normal distribution. By default, we simulate a target distribution with 5 explanatory variables and add a Gaussian noise with a standard deviation of 0.5.

Task	
Number of Explanatory Variables	5
Initial Coefficient (β_i) Weights	Standard Normal Distribution
Explanatory Variable (x_i)	Standard Normal Distribution
Noise (ϵ)	Normal Distribution $\mu = 0, \sigma = 0.5$

Table 3.1: Task Summary

Next, following the idea that human decisions tend to be linear, we represent our learning agent as a linear regression algorithm based on gradient descent. The agent's objective is to approximate a target function $y = \beta_0 + \beta_1 x_1 + ... + \beta_n x_n$ by using samples (x, y). That is, given a set of factors x and the outcome y, the agent tries to learn the coefficients β. Using a training set of size n, The algorithm learns by minimizing the mean squared error cost function J:

$$J = \frac{1}{n} \sum_{i=0}^{n} (f(x_i) - h(x_i))^2. \tag{3.4}$$

The cost function measures the difference between the true target value $f(x)$ and the value predicted by an agent's hypothesis $h(x)$. To minimize the cost function, the agent uses gradient descent, an iterative optimization algorithm that finds the local minimum of a differentiable function. To find the coefficients β that minimize the cost function, the algorithm calculates the partial derivative of the cost function with respect to β:

$$\frac{dJ}{d\beta} = \frac{-2}{n}\sum_{i=0}^{n} x_i(f(x_i) - h(x_i)). \tag{3.5}$$

The derivative of a function represents the rate of change of one variable in relation to another at a given point on a function. In each iteration of the gradient descent, the algorithm updates the weights by stepping down the cost function in the direction of the steepest descent. The size of each step is determined by the learning rate α. In each iteration, the weight is updated by,

$$\beta = \beta - \alpha * \frac{dJ}{d\beta}. \tag{3.6}$$

By default, we select the agents' initial coefficient weights from a normal distribution, set the learning rate to 0.05, and perform the gradient descent updates for 100 iterations.

Learning Agent	
Initial Coefficient Weights	Standard Normal Distribution
Learning Rate (α)	0.05
Gradient Descent Iterations	100

Table 3.2: Learning Agent Summary

3.3 Measuring Polarization

Now that we have defined the task and the learning agent, let us instantiate the problem and discuss how we can measure polarization. Consider an example in which the learning agents are doctors forming beliefs about the safety of a newly developed vaccine. Two doctors form beliefs about the vaccine by observing their own set of patients, along with the outcomes of the treatments. The patients that the doctors observe represent the training set. For a particular training sample (x, y), x is a vector of values that describe the treatment such as the patient's age, the patient's blood pressure, the dosage of the vaccine, and so on. Then, the label y represents the patient's response to the vaccine. As aforementioned, we represent the beliefs as being distributed on a normalized spectrum along one dimension. For instance, a large y indicates that the treatment was effective and safe.

After the doctors form their beliefs, they are both given the same test set of new patients and are told to predict how they will respond to the vaccine. We measure polarization based on their predictions on the test set. We propose three different ways of measuring polarization: variance, continuous disagreement, and discrete disagreement.

3.3.1 Variance

In machine learning, bias-variance decomposition is a way of analyzing a learning agent's expected error with respect to a particular problem as a sum of three terms: bias, variance, and irreducible noise.

$$Expected\ Error = bias^2 + variance + noise. \tag{3.7}$$

The squared error, $err(h) = (h(x) - f(x))^2$, can be decomposed into bias and variance as follows [7]:

$$bias^2 = [E[(h(x)] - f(x)]^2 \tag{3.8}$$

$$variance = E[[h(x) - E[h(x)]]^2] \tag{3.9}$$

The bias represents the extent to which the average prediction over all data sets differs from the target function, and the variance measures the extent to which the predictions for individual data sets vary around their average. In the context of belief formation, the bias measures how far the agents' beliefs are from the objective truth, and the variance measures how far the agents' beliefs are from each other. In the context of belief polarization, we are particularly interested in the variance term. Since variance measures the spread of the beliefs, we say that a large variance means greater polarization.

3.3.2 Continuous Disagreement

Conceptually, we can say that two agents are polarized when their predictions on the same test set disagree significantly. If two doctors make many conflicting predictions on the same set of new patients, they must have a polarizing belief system. We define the *continuous disagreement*, $\Delta_{continuous}$, as the average difference between the two agents' predictions on a test set. Given a test set of size m, the continuous disagreement between two agents h_1 and h_2 is

$$\Delta_{continuous} = \frac{1}{m} \sum_{i=0}^{m} |(h_1(x_i) - h_2(x_i))|. \tag{3.10}$$

The continuous disagreement measures the average difference in two agents' predictions on the entire test set. Similar to the variance, continuous disagreement measures the spread of agents' beliefs. A large continuous disagreement suggests that opinions are far apart in content.

3.3.3 Discrete Disagreement

Rather than interpreting beliefs as a point in a belief spectrum, we may want to interpret them as discrete choices. To do so, we map y into discrete classes based on a threshold. We set a threshold t around 0, and convert y based on the following intervals:

$$y = \begin{cases} -1 & y < -t \\ 0 & -t \leq y \leq t \\ 1 & t < y \end{cases} \tag{3.11}$$

For example, the labels can be interpreted as the doctors' decisions: -1 represents a decision not to vaccinate the patient, 0 represents an inconclusive decision, and 1 represents a decision to vaccinate the patient. We define the *discrete disagreement*, $\Delta_{discrete}$, as the average difference between the two agents' discrete decisions on a test set. For our model, we set the threshold $t = 0$. Given a test set of m, the discrete disagreement between two agents is

$$\Delta_{discrete} = \frac{1}{m} \sum_{i=0}^{m} I(h_1(x_i)) \neq (h_2(x_i))) \tag{3.12}$$

where I is an indicator function, which evaluates to 1 if the expression is true, and 0 otherwise. Threshold models are common in cognitive science. For instance, threshold values in practice could represent a linear combination of a set of neural firing rates, at which the accumulation of evidence triggers an action [33]. The model formulates a setting where the agent only makes a decision for either choice A or B when the decision variable reaches a certain threshold. High discrete disagreement suggests that agents make conflicting decisions on many test samples.

Chapter 4

Simulations

4.1 Simulation Description

The goal of the simulation is to observe how sample size affects polarization. We create groups of 100 learning agents based on the number of training samples they observe. We simulate sample sizes from 2 to 20. For instance, the first group consists of 100 agents who only observe 2 samples, the second group consists of 100 agents who only observe 3 samples, and so on. After the agents learn their belief functions, they predict the same test set consisting of 100 random test samples from the target distribution. For each group, we measure polarization based on the predictions on the test set. The target distribution and the learning agents are initialized using the default hyperparameters described in the previous chapter.

4.2 Visualization of Polarization

Before we calculate the different measures of polarization, let us visualize what a high polarization looks like compared to a low polarization.

Figure 4.1 shows each group's predictions on a single random test sample. The

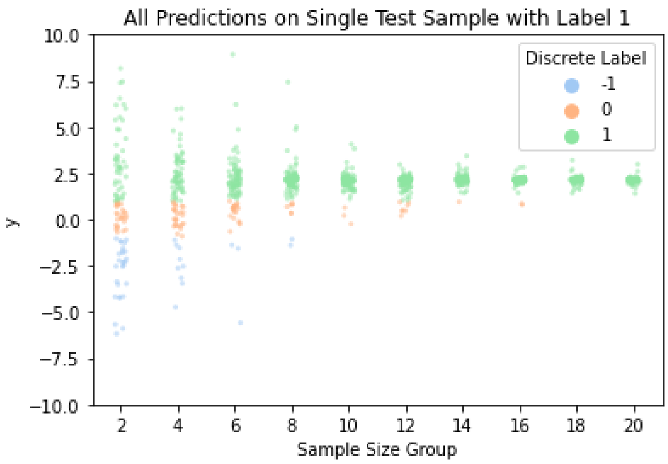

Figure 4.1: Prediction Distribution on Single Test Sample

chosen test sample has a target value of approximately 2.0 (or a discrete label of 1). When agents rely on a few samples, the predictions have a large spread. For instance, the difference between the highest and the lowest value is greater in groups where agents observe fewer samples. The spread indicates how far apart the two extreme opinions are in content. As agents observe more samples, the predictions eventually converge to the target value. When we consider the beliefs as discrete choices, the figure shows that agents who observe small samples predict conflicting labels. Even though the target label is 1, many agents predict 0 or -1. When agents observe more samples, however, they all predict the correct label.

To visualize a large polarization between two agents on a test set, consider figure 4.2. The first column of the graph shows 100 test samples sorted based on the target values for visualization. The two columns to the right display predictions by two agents who observed only 2 train samples. There is a clear disparity between the two agents' predictions on the test samples. For instance, agent A predicts positive values for many samples that agent B predicts negative values. That is, given the same set of patients, one doctor believes vaccination will be risky, while the other doctor believes vaccination will be safe. On the other hand, figure 4.3

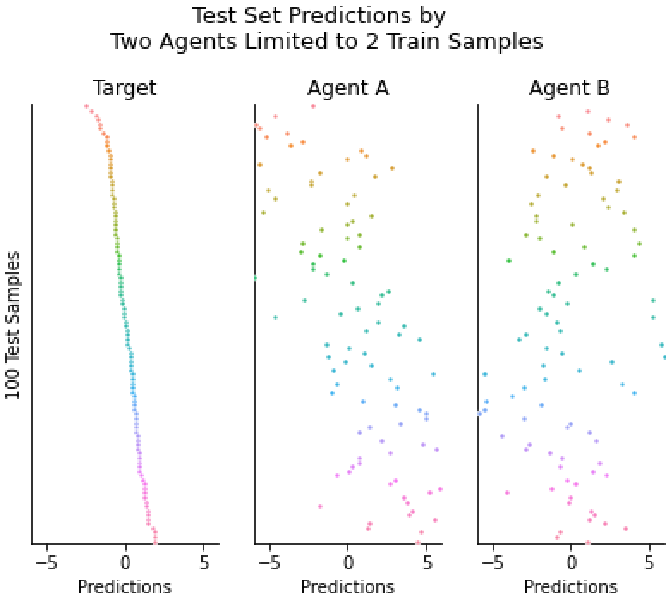

Figure 4.2: Test Set Predictions by Two Agents Limited to 2 Train Samples

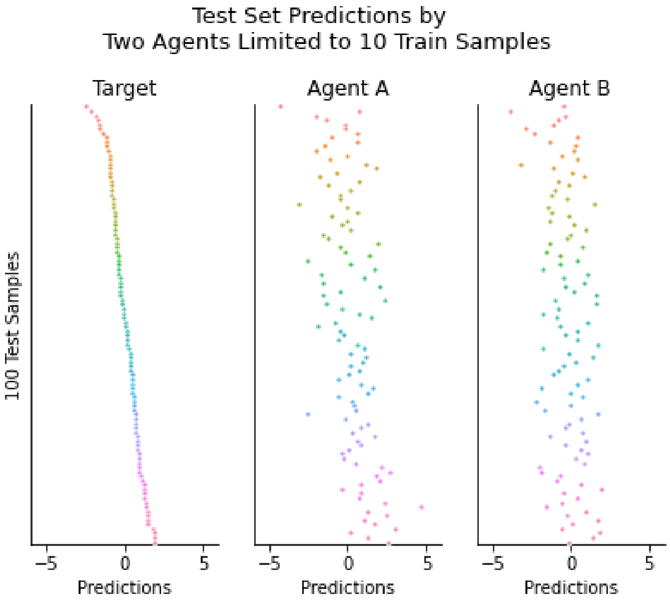

Figure 4.3: Test Set Predictions by Two Agents Limited to 10 Train Samples

shows the predictions by two agents who observe more train samples. When the agents observe a larger number of samples, the two agents' predictions are more symmetrical to each other and to the target.

4.3 Bias and Variance Analysis

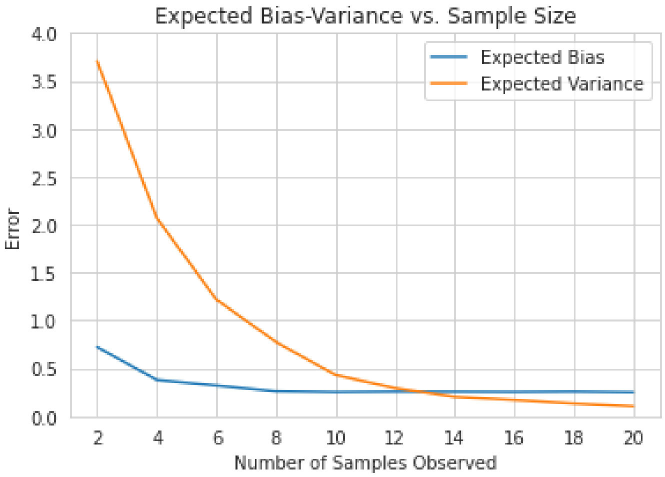

Figure 4.4: Expected Bias-Variance vs. Sample Size

Now, we measure polarization using the variance. Figure 4.4 shows the expected bias and the expected variance for agents who observe a different number of training samples. The result shows that as agents observe more samples, their expected variance decreases monotonically. This suggests that when two people observe more information, their beliefs tend to be more aligned with each other. For example, two doctors who have observed more previous patients are expected to make more similar predictions about new patients. The figure also shows that when agents observe a sufficient amount of samples, the expected variance is minimized and additional samples do not reduce the variance. This suggests that once people observe enough information, their beliefs converge.

Similarly, the expected bias initially decreases as agents observe more samples. However, the expected bias is notably smaller and plateaus faster than the expected variance. In our model, the variance contributes much more to the total error than the bias when agents rely on small samples. Note that the reason the agents have low bias is that linear agents are suited for solving linear regression tasks. Meaning, our model describes a setting in which people are learning about a topic that they can fully comprehend once they observe enough information. Suppose our agents are given many tasks and have to learn many target distributions, but samples are costly. If the agent's goal is to be as accurate as possible across many tasks, but samples are costly, they may choose to only observe a few samples. This is because few samples are sufficient to minimize the expected bias. However, small samples lead to high variance, which results in greater polarization.

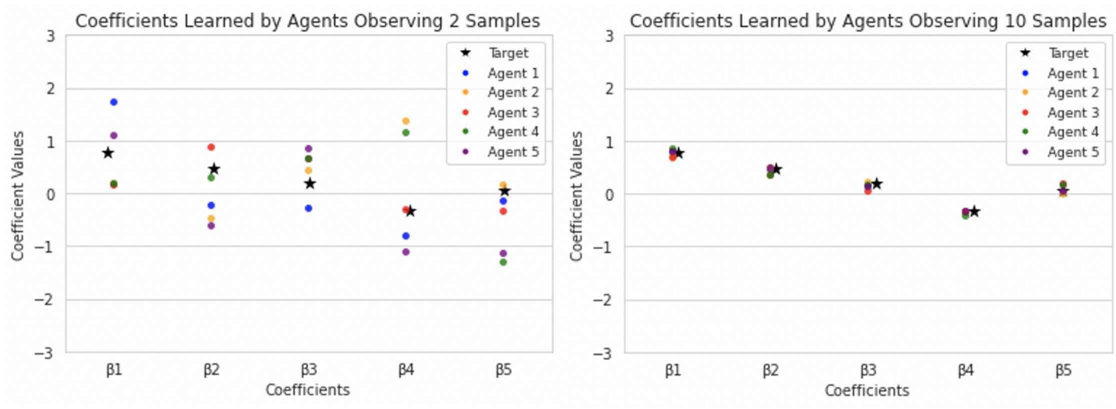

Figure 4.5: Learned Coefficients Based on Number of Samples Observed

The variance in agents' predictions stems from how they learn the coefficients of the explanatory variables. Consider an explanatory variable that represents a patient's blood pressure. When two doctors only observe a few patients, they may discern a different understanding of how blood pressure affects the outcome of the treatment. Consider figure 4.5. The graph on the left shows the coefficient weights learned by five random agents who only observed 2 samples. The graph on the right shows coefficient weights learned by five random agents who observed 10 samples. Even though the agents learn using samples drawn from the same target distribution, small sample reliance leads to a large dispersion in the learned weights. However, once the agents observe sufficient samples, their weights align with the target weights and with one another.

4.4 Disagreement Simulations

Next, we measure polarization using the two disagreement measures we proposed: continuous disagreement and discrete disagreement. We repeat the simulation by varying three hyperparameters: the number of explanatory variables, the noise, and the number of gradient descent iterations. When we vary one hyperparameter value, we fix the rest of the hyperparameters to the default value.

4.4.1 Explanatory Variables

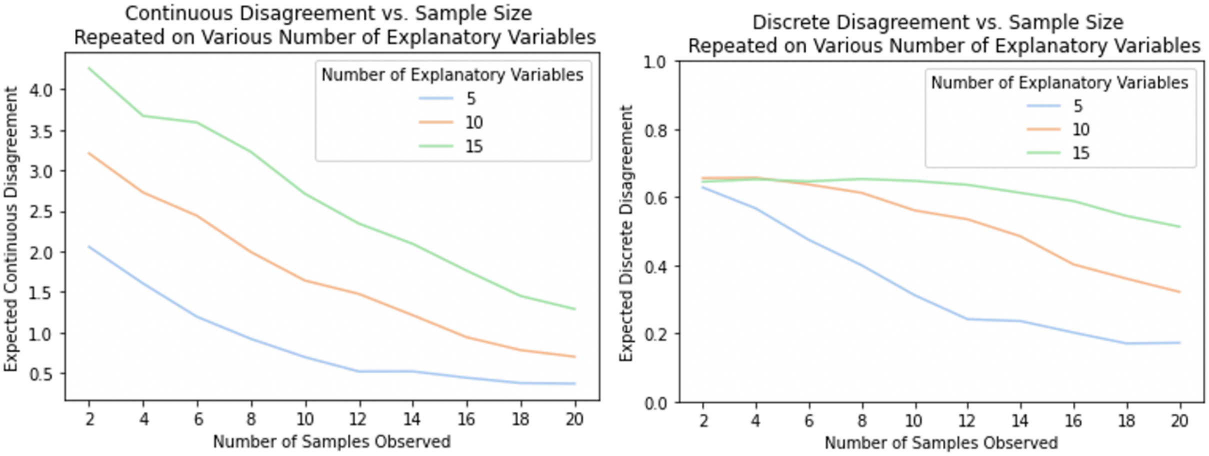

Figure 4.6: Disagreement vs. Sample Size: Varying Number of Explanatory Variables

Figure 4.6 shows the expected disagreement between two agents, repeated on a varying number of explanatory variables. The graph on the left shows the expected continuous disagreement and the graph on the right shows the expected discrete disagreement.

First, consider the continuous disagreement. Similar to the variance, the expected continuous disagreement decreases monotonically as agents observe more samples. In addition, the expected continuous disagreement is higher on tasks with a greater number of explanatory variables. More explanatory variables imply that the task is more complex. For instance, a subject like vaccination requires one to consider more factors than choosing to turn left or right in traffic. The graph indicates that people tend to disagree more strongly on topics that are more complex. Assuming people observe the same sample size, the difference between two people's beliefs is expected to be larger on more complex tasks.

Next, consider the discrete disagreement, where the response variables are mapped into discrete classes. When agents observe a few samples, their expected discrete disagreement is approximately 2/3. Meaning, two agents will on average disagree on 2/3 of the decisions on new test samples. This indicates that when people make discrete decisions based on a few pieces of information, their decisions are almost random. The expected discrete disagreement tends to decrease as agents observe more samples. However, when agents deal with complex tasks with many variables, the disagreement does not decrease until they have observed a sufficient number of samples. This suggests that people may need to acquire sufficient information before they make a decision. When the subject of interest is complex, people need to consider more information before they are "convinced" to make a decision.

4.4.2 Noise

Figure 4.7 shows the simulation repeated on varying levels of sample noise. We represent noise as any general notion of a noise that disturbs the quality of the information. For instance, noise may stem from a poor description of the information or inefficient delivery. Based on the simulation results, we see that as noise increases, both the expected continuous disagreement and the expected discrete disagreement

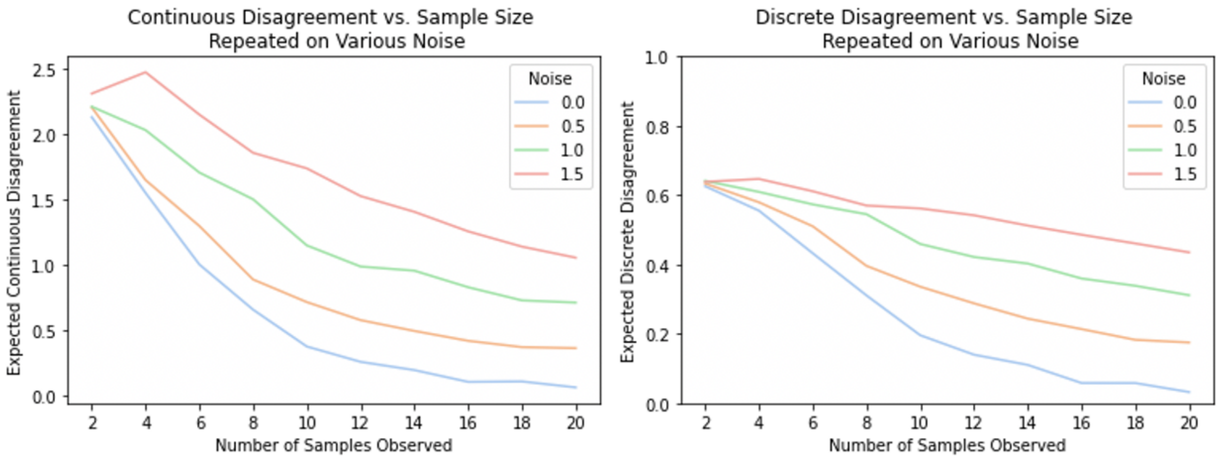

Figure 4.7: Disagreement vs. Sample Size: Varying Noise

between two agents increase. Even when the agents observe sufficient samples, the disagreement remains high when the noise is high. Meaning, when the quality of the information is low, people can still have high polarization even after observing a large amount of information.

4.4.3 Gradient Descent Updates

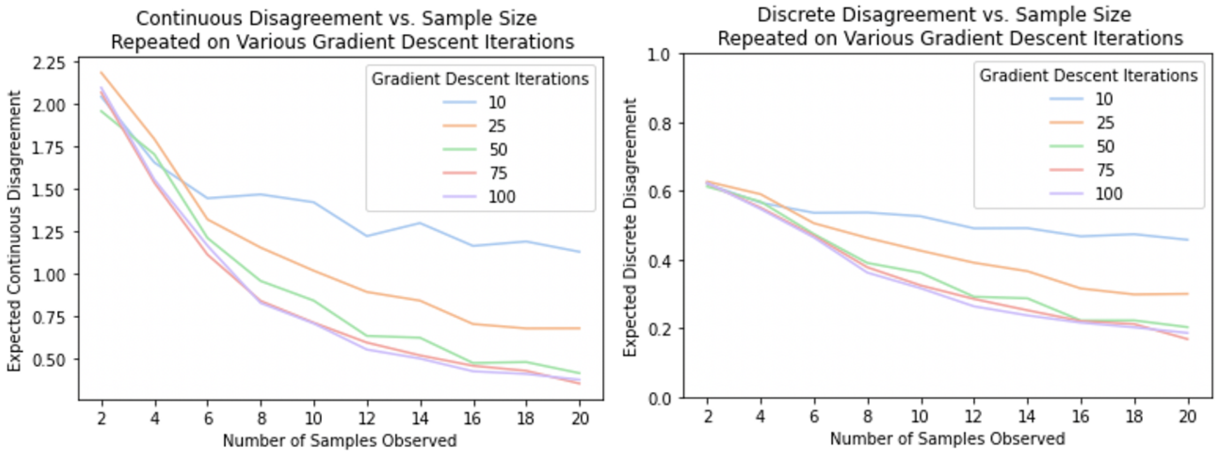

Figure 4.8: Disagreement vs. Sample Size: Varying Gradient Descent Updates

Figure 4.8 shows the simulation repeated on varying numbers of gradient descent iterations that the learning agents perform. The number of iterations represents the effort that people put to process the information. Few gradient descent updates indicate that people put little effort into analyzing the information they are given. People may put in little effort due to cognitive constraints, similar to why people rely on small samples. Based on the simulation results, both the expected continuous disagreement and the expected discrete disagreement decrease when agents perform more gradient descent updates. Once the agents perform enough gradient descent updates, their beliefs converge. When agents only perform a few updates, their disagreement remains high even when they observe sufficient samples. Meaning, if people do not fully process the information they are given, they can polarize despite receiving a large amount of information.

Chapter 5

Conclusion

5.1 Summary

We questioned why people form polarizing beliefs and proposed a new explanation. In an increasingly connected world, people are exposed to an abundance of information concerning a multitude of subjects, but processing information is costly. We discussed that in order to maximize the expected reward over time, people may rely on a few samples to form beliefs. However, small samples are highly variable and may lead to polarization. To test this hypothesis, we explored a new method of modeling polarization based on an empirical analysis of machine learning. After designing a machine learning framework that emulates human learning, we used simulations to examine the impact of small sample reliance on belief polarization. Our simulation results confirmed our hypothesis that small sample reliance leads to greater polarization. The results also indicated that complex problems, high levels of noise, and low cognitive efforts also lead to greater polarization.

5.2 Practical Application

Based on our findings, we offer possible suggestions for mitigating polarization in the real world. Our simulations indicate that polarization decreases when agents observe more samples, assuming that people observe information from similar sources. However, we also found that when the problem is too complex, has high noise, or is difficult to process, increasing the sample size only reduces polarization marginally. Therefore, we suggest exposing people to more high-quality samples that can be easily digested at little cost. The first challenge is making sure that the information people observe is of high quality. Many studies have explored content-selection algorithms to recommend samples that accurately represent the population [28]. For instance, one algorithm [20] attempts to reduce polarization by deterring like-minded people from creating echo chambers. The second challenge is exposing people to more samples. Since people rely on small samples due to their cost, we suggest methods of delivering information in a cost-efficient manner. For instance, consider people browsing news articles online. Typically, people only read one or two articles regarding some topic before moving on to the next topic. That is, people are unlikely to research further and look for various perspectives. We speculate that if people are presented with multiple headlines on the same topic from diverse sources, people may be more inclined to consider various perspectives. In addition, we may also provide readers with well-curated summaries or abstracts of the topic that make the key points easier to absorb. Assuming people have a fixed amount of time and energy to consider information, making the information cost-efficient to understand may be a promising strategy.

5.3 Future Work

This thesis was an experimental study exploring a novel approach to modeling human cognition and polarization. Given the countless configurations involved in the study of belief formation and the social phenomenon of polarization, we believe a flexible framework like machine learning is appropriate and can provide valuable insights. For instance, our work only explores one of many possible ways in which we can instantiate the learning agent and the learning task. Based on the knowledge that human decisions tend to exhibit linear properties, we expressed the learning agent as a linear model. However, one may explore a different, perhaps non-linear, learning algorithm to depict how humans learn. Similarly, based on the knowledge that many relationships in the world are linear, we expressed the task as a linear problem. However, one may design a unique machine learning task depending on the focus of their study. In addition, our model assumed a supervised learning framework. However, different frameworks such as unsupervised learning, reinforcement learning, or even deep learning may serve as an appropriate expression of how humans learn. As the field of machine learning grows, how we can utilize machine learning to better understand cognition will be a topic of great interest.

Bibliography

[1] Norman H Anderson. Foundations of information integration theory. 1981.

[2] Charles Angelucci and Andrea Prat. Is journalistic truth dead? measuring how informed voters are about political news. *Measuring How Informed Voters Are about Political News (June 30, 2021)*, 2021.

[3] Marina Azzimonti and Marcos Fernandes. Social media networks, fake news, and polarization. *European Journal of Political Economy*, page 102256, 2022.

[4] Eytan Bakshy, Solomon Messing, and Lada A Adamic. Exposure to ideologically diverse news and opinion on facebook. *Science*, 348(6239):1130–1132, 2015.

[5] Claude Bélisle and James Cresswell. The effects of a limited memory capacity on foraging behavior. *Theoretical Population Biology*, 52(1):78–90, 1997.

[6] Yoella Bereby-Meyer and Ido Erev. On learning to become a successful loser: A comparison of alternative abstractions of learning processes in the loss domain. *Journal of mathematical psychology*, 42(2-3):266–286, 1998.

[7] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

[8] Renee Bowen, Danil Dmitriev, and Simone Galperti. Learning from shared news: when abundant information leads to belief polarization. Technical report, National Bureau of Economic Research, 2021.

[9] Aaron Bramson, Patrick Grim, Daniel J Singer, William J Berger, Graham Sack, Steven Fisher, Carissa Flocken, and Bennett Holman. Understanding polarization: Meanings, measures, and model evaluation. *Philosophy of science*, 84(1):115–159, 2017.

[10] Colin Camerer. General conditions for the success of bootstrapping models. *Organizational Behavior and Human Performance*, 27(3):411–422, 1981.

[11] Arun G Chandrasekhar, Horacio Larreguy, and Juan Pablo Xandri. Testing models of social learning on networks: Evidence from two experiments. *Econometrica*, 88(1):1–32, 2020.

[12] Robyn M Dawes, David Faust, and Paul E Meehl. Clinical versus actuarial judgment. *Science*, 243(4899):1668–1674, 1989.

[13] Carroll Doherty, Jocelyn Kiley, and Nida Asheer. Partisan antipathy: More intense, more personal. *Pew Research Center*, 2019.

[14] Jeffrey L Elman. Learning and development in neural networks: The importance of starting small. *Cognition*, 48(1):71–99, 1993.

[15] Joan Esteban and Gerald Schneider. Polarization and conflict: Theoretical and empirical issues: Introduction. *Journal of Peace Research*, 45(2):131–141, 2008.

[16] Roland G Fryer Jr, Philipp Harms, and Matthew O Jackson. Updating beliefs when evidence is open to interpretation: Implications for bias and polarization. *Journal of the European Economic Association*, 17(5):1470–1501, 2019.

[17] Federico Germani and Nikola Biller-Andorno. The anti-vaccination infodemic on social media: A behavioral analysis. *PloS one*, 16(3):e0247642, 2021.

[18] Benjamin Golub and Matthew O Jackson. How homophily affects the speed of learning and best-response dynamics. *The Quarterly Journal of Economics*, 127(3):1287–1338, 2012.

[19] Nika Haghtalab, Matthew O Jackson, and Ariel D Procaccia. Belief polarization in a complex world: A learning theory perspective. *Proceedings of the National Academy of Sciences*, 118(19), 2021.

[20] Mathew D Hardy, Bill D Thompson, PM Krafft, and Thomas L Griffiths. Bias amplification in experimental social networks is reduced by resampling. *arXiv preprint arXiv:2208.07261*, 2022.

[21] Reid Hastie and Robyn M Dawes. *Rational choice in an uncertain world: The psychology of judgment and decision making.* Sage Publications, 2009.

[22] Robin Hau, Timothy J Pleskac, and Ralph Hertwig. Decisions from experience and statistical probabilities: Why they trigger different choices than a priori probabilities. *Journal of Behavioral Decision Making*, 23(1):48–68, 2010.

[23] Robin Hau, Timothy J Pleskac, Jürgen Kiefer, and Ralph Hertwig. The description–experience gap in risky choice: The role of sample size and experienced probabilities. *Journal of Behavioral Decision Making*, 21(5):493–518, 2008.

[24] Bernd Heinrich. Resource heterogeneity and patterns of movement in foraging bumblebees. *Oecologia*, 40(3):235–245, 1979.

[25] Leah Henderson and Alexander Gebharter. The role of source reliability in belief polarisation. *Synthese*, 199(3):10253–10276, 2021.

[26] Yaakov Kareev. Seven (indeed, plus or minus two) and the detection of correlations. *Psychological review*, 107(2):397, 2000.

[27] Ralph L Keeney, Howard Raiffa, and Richard F Meyer. *Decisions with multiple objectives: preferences and value trade-offs*. Cambridge university press, 1993.

[28] David Lazer. The rise of the social algorithm. *Science*, 348(6239):1090–1091, 2015.

[29] Falk Lieder and Thomas L Griffiths. Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and brain sciences*, 43, 2020.

[30] Charles G Lord, Lee Ross, and Mark R Lepper. Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of personality and social psychology*, 37(11):2098, 1979.

[31] Miller McPherson, Lynn Smith-Lovin, and James M Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*, 27(1):415–444, 2001.

[32] George A Miller. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review*, 63(2):81, 1956.

[33] Paul Miller. *Decision Making, Threshold*, pages 1–4. Springer New York, New York, NY, 2013.

[34] Amy Mitchell, Jeffrey Gottfried, Jocelyn Kiley, and Katerina Eva Matsa. Political polarization & media habits. 2014.

[35] Iris Nevo and Ido Erev. On surprise, change, and the effect of recent outcomes. *Frontiers in psychology*, 3:24, 2012.

[36] Brendan O'Connor. Biased evidence assimilation under bounded bayesian rationality. *Unpublished master's thesis, Stanford University*, 2006.

[37] Laura Santhanam. Why americans have grown more hesitant about the covid-19 vaccine. *PBS News Hour*, 2020.

[38] Fernando P Santos, Yphtach Lelkes, and Simon A Levin. Link recommendation algorithms and dynamics of polarization in online social networks. *Proceedings of the National Academy of Sciences*, 118(50), 2021.

[39] Leonard J Savage. *The foundations of statistics*. Courier Corporation, 1972.

[40] Amos Tversky and Daniel Kahneman. Belief in the law of small numbers. *Psychological bulletin*, 76(2):105, 1971.

[41] Leslie G Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

[42] John Von Neumann and Oskar Morgenstern. Theory of games and economic behavior princeton. *Princeton University Press*, 1947:1953, 1944.

[43] Edward Vul, Noah Goodman, Thomas L Griffiths, and Joshua B Tenenbaum. One and done? optimal decisions from very few samples. *Cognitive science*, 38(4):599–637, 2014.