# Belief polarization in a complex world: A learning theory perspective

Nika Haghtalab[a], Matthew O. Jackson[b,c,1], and Ariel D. Procaccia[d]

[a]Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA 94720; [b]Department of Economics, Stanford University, Stanford, CA 94305; [c]Santa Fe Institute, Santa Fe, NM 87501; and [d]School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138

Contributed by Matthew O. Jackson, March 18, 2021 (sent for review May 20, 2020; reviewed by Aislinn Bohren, Grant Schoenebeck, and Omer Tamuz)

**We present two models of how people form beliefs that are based on machine learning theory. We illustrate how these models give insight into observed human phenomena by showing how polarized beliefs can arise even when people are exposed to almost identical sources of information. In our first model, people form beliefs that are deterministic functions that best fit their past data (training sets). In that model, their inability to form probabilistic beliefs can lead people to have opposing views even if their data are drawn from distributions that only slightly disagree. In the second model, people pay a cost that is increasing in the complexity of the function that represents their beliefs. In this second model, even with large training sets drawn from exactly the same distribution, agents can disagree substantially because they simplify the world along different dimensions. We discuss what these models of belief formation suggest for improving people's accuracy and agreement.**

belief polarization | learning theory

ECONOMIC SCIENCES

COMPUTER SCIENCES

In 1998, *The Lancet*, a medical journal, published a study linking the MMR (measles, mumps, and rubella) vaccine to autism. Although this study has since been retracted and its results refuted, it is still a rallying cry for the modern antivaccination movement. Periodic outbreaks of measles are often associated with pockets of resistance to vaccination—like the measles public health emergency in New York City in April 2019 (1). Research shows that even though the opinions of Americans about vaccines are highly polarized, they are not divided along the usual political fault lines that typically correspond to disparate sources of information. Instead, the more political a person is (in either direction), the more likely one is to think vaccines are unsafe (2).

Divergence in beliefs about vaccines is just one example of belief polarization, which of course, is not a new phenomenon. Such polarization is sometimes attributed to people's exposure to divergent sources of information—"echo chambers"—which can exist because people tend to interact with others who share their background as well as opinions (3, 4). Cable television, the internet, social media, and other technologically mediated communication can immerse users in content that is tailored to their existing preferences and shared by like-minded people (5–7). In that reality, polarization is a natural outcome; when different people are exposed to significantly different sources of information, they can arrive at different conclusions.

However, there is substantial evidence that polarization also arises even when agents are exposed to the same source of information (8–11). A long-standing explanation for this is a tendency of people to interpret information to confirm what they already believe (8, 9, 12, 13). Here, we provide a different explanation for this phenomenon based on foundational models of how people learn. We build two such models that are based on machine learning theory and show that they each can lead to situations in which people's beliefs differ substantially, even when faced with almost identical information.

The models that we propose differ from the standard models of how people form beliefs. A pair of landmark theorems by von Neumann and Morgenstern (14) and Savage (15) provided a foundation of rational decision making as a Bayesian exercise. Unimpeded, full rationality has become synonymous with having a prior distribution over possible consequences from different actions and processing any information by updating that prior distribution according to Bayes' rule. Although one can rationalize almost any pattern of beliefs and behaviors as being Bayesian (16), the priors have to become increasingly convoluted to explain why whole societies remain systematically polarized when confronted repeatedly with the same information. Moreover, people's behavior has failed to exhibit Bayesian updating along basic dimensions (17). This has led to a variety of models of bounded rationality or limited observability in which beliefs are updated by some other adaptive or reinforcement manner (18–23), are based on some misspecification or misunderstanding (24, 25), or are derived from observations of the actions of others who may have different preferences (26, 27).

Paradoxically, researchers model people as learning by updating beliefs via Bayes' rule or some boundedly rational process, but then, researchers themselves learn in different ways; they build models of the world and then discard those models for new ones when a model no longer sufficiently matches available data. Methods from regression analysis, nonparametric statistics, and machine learning involve changing the parameters and even the basic structure of a model as new data become available, and sometimes researchers even invent a new class of models when old ones no longer perform well. That behavior corresponds more closely to human learning, which can involve coming to a whole new understanding of how the world works after going through a novel experience (28). It does not correspond to a

## Significance

**Differences in beliefs within a society are a prevalent human phenomenon. A standard explanation for polarized beliefs relies on "echo chambers" that expose people to different sources of information. However, there is ample evidence that people sustain different beliefs even when faced with the same information, and they interpret that information differently—facts often attributed to a confirmatory bias. We suggest models of how humans form beliefs based on machine learning theory. These models show how stark differences in beliefs can arise and endure due to human limitations in interpreting complex information. Our framework illuminates inherent challenges and potential ways of overcoming polarization.**

Author contributions: N.H, M.O.J., and A.D.P. designed research, performed research, and wrote the paper.

Reviewers: A.B., University of Pennsylvania; G.S., University of Michigan; and O.T., California Institute of Technology.

The authors declare no competing interest.

[1] To whom correspondence may be addressed. Email: jacksonm@stanford.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2010144118/-/DCSupplemental.

Published May 3, 2021.

person who preconceived all possible models and updated a prior belief over those models as new information became available. These observations suggest exploring a model of humans as machine learners: forming belief functions that can completely change with experience. Such a model can give alternative explanations for polarization and lead to policy insights.

Our machine learning-based approach can be seen as being in line with Simon's (29) description of human decision making: listing alternatives and assessing each of their consequences based on past experiences and other available information. It is also more in line with a view of bounded rationality as involving some complexity costs (30, 31). It differs from modeling people as finite automata—which has been used in the study of repeated games (32, 33)—in that it provides a paradigm in which the natural objects are belief functions that make predictions based on past data, and thus, it offers a direct way of modeling belief formation.

Below, we provide two models that each generate differences in the beliefs of two people being exposed to nearly the same data. In the first, two people each start with their own past observations that consist of past circumstances and accompanying outcomes, and those two datasets can be perfectly fit using two different functions that map circumstances into outcomes. However, these observations are subjective in that they disagree for some circumstances. When they share their past observations with each other, there is no longer any (deterministic) function that fits the combined datasets. If the two people do not completely share their observations and they are each slightly biased toward having more of their own observations, then they can end up having very different optimal predictions of outcomes for the same circumstances.

The second model brings complexity costs into play and is based on a purely objective reality that everyone would completely agree upon with enough information and no constraints on forming beliefs. Instead, in this model the two people each face a cost of the complexity of their model. As an example, suppose that 10 dimensions of the circumstance matter roughly equally in determining an outcome but that it becomes prohibitively expensive for a person to build a mental model that tracks more than 7 of those dimensions. Different people who have seen even slightly different samples of circumstances and outcomes can end up finding different sets of seven dimensions being most effective at explaining what they have seen. For instance, one person may have seen more circumstances for which scientific evidence is an important variable that helps explain the outcome, while another individual might have seen more circumstances for which politics are an important predictor of the outcome. These two can end up paying attention to different dimensions when faced with making a prediction for some circumstances and for example, end up with very different beliefs about climate change. Slight differences in samples can lead two people to form significantly different belief functions concerning the same phenomenon.

After presenting the models and results, we discuss their implications for improving the accuracy and agreement of people's beliefs.

## An Overview of Our Approach and Results

To model beliefs, we draw on the discriminative "probably approximately correct" learning framework of machine learning theory (34).

An agent has seen past data, referred to as the training set, in the form of instances from a set $\mathcal{X}$ that describe possible circumstances, together with a label or an outcome from a set $\mathcal{Y}$ for each instance. The random draw of the agent's training set (instances together with their labels/outcomes) is represented by a distribution $\mathcal{D}$ on $\mathcal{X} \times \mathcal{Y}$, which describes the samples of data that the agent is likely to see. Based on the training set, the agent develops a model in the form of a function $f : \mathcal{X} \to \mathcal{Y}$ that maps the input space to the set of labels. This is done by choosing (from a prespecified class of functions) the one that performs best according to some objective function that captures how well the function matches the training set, possibly adjusted for a cost for how complex the function is. We refer to $f$ as a belief function.

Let us instantiate the terminology. Consider an example in which the agent is a doctor forming beliefs about the effectiveness of different medical treatments. There, an instance $x \in \mathcal{X}$ consists of a patient with a set of symptoms and history together with a treatment that is given (e.g., a 50-y-old male smoker with hypertension who is given a specific drug to reduce his blood pressure), and the $y$ would be the doctor's perception of the outcome of the treatment (the blood pressure went down by a certain amount). These observations could be based on personal experiences of the doctor with her own patients, from discussions with other doctors, or from reading medical journals. The belief function $f(x) = y$ then describes what the doctor best predicts would be the outcome for a patient with the set of symptoms, history, and treatment described by $x$. Different doctors will have seen different training sets—different observations of pairs $(x, y)$—over their careers and so, may have arrived at different functions $f$. This example also applies to the effects of medical treatments that have been more polarizing, such as whether a vaccine would cause autism when given to a child of a certain age. It also applies to scenarios where a parent, who is interested in vaccinating their child, forms a decision based on scientific studies and anecdotal experiences of other vaccinated children.

In this formulation, people's beliefs are represented by deterministic functions. These functions can be thought of as expressions of a person's opinion or plan of action for any situation they may face. Expressing beliefs instead as probabilities or distributions requires an accurate expression of probabilistic events that are known to be difficult for people. By contrast, deterministic functions are more easily interpreted and explained by humans. Moreover, modeling beliefs as functions focuses on the formation of opinion about circumstances, while belief distributions are further complicated by the need to express the likelihood of facing circumstances.

To develop intuition for our results, suppose that two agents see training sets drawn from the same distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$ and that this distribution is realizable [i.e., there is a deterministic belief function $f^*$ such that $f^*(x) = y$ for every $(x, y)$ in the support of $\mathcal{D}$]. In this classic setting, without any limitations on the complexity of their beliefs and training sets that grow without bound, agents who form beliefs to best fit the data will eventually learn belief functions that almost entirely agree with $f^*$ and hence, with each other on the support of $\mathcal{D}$.

Our two models deviate from this basic setting in fundamentally different ways.

In our first model—the mixed subjective model—agents see different, yet highly overlapping, labels; in this sense, the agents each have their own "subjective" views based on differing personal histories and perspectives. In our medical example, agents have their own evaluations of treatments based on each agent's experience. To formalize this, suppose that two agents are associated with two different realizable distributions $\mathcal{D}_1$ and $\mathcal{D}_2$ over $\mathcal{X} \times \mathcal{Y}$, which have the same marginal over $\mathcal{X}$. Learning from these two different distributions can naturally give rise to highly polarized belief functions.

Now suppose that, perhaps with the aim of finding common ground, the two agents share some of their training sets, leading them both to observe biased mixtures of the two distributions. Specifically, let agent 1 train on data drawn from

$(1/2+\varepsilon)\mathcal{D}_1+(1/2-\varepsilon)\mathcal{D}_2$, while agent 2 trains on data from $(1/2-\varepsilon)\mathcal{D}_1+(1/2+\varepsilon)\mathcal{D}_2$, for some small $\varepsilon>0$. We show that, even when mixing to almost even proportions, the two agents will still learn substantially different belief functions. After interacting, even in the extreme case in which two agents give almost equal weight to the two sources, if each has slightly more examples from their own source, then they could end up with very different best-fitting views.

In our second model—the complex objective model—we go back to the "objective" setting in which the distribution from which agents' training sets are drawn is the same: an objective truth. Moreover, let this be a realizable distribution $\mathcal{D}$ over $\mathcal{X}\times\mathcal{Y}$ so that there is an associated deterministic $f$ such that $f(x)=y$ for every $x, y$ in the support of $\mathcal{D}$. So, for example, with sufficient research and knowledge, agents would agree and be correct in a deterministic prediction. In this model, we consider agents who pay a cost for the complexity of the function $f$ that they fit. Agents' belief functions are those that minimize a weighted average of the number of errors they make and the cost of the complexity of the function (similar to a Lagrangian expression of a Lasso [least absolute shrinkage and selection operator] regression).

The complex objective model gives rise to realizable distributions $\mathcal{D}$, such that having a nontrivial cost of complexity leads agents to learn belief functions that result in substantial disagreement, even as the number of observed examples goes to infinity. The basic structure of such instances is that they have many dimensions or variables that actually matter, more or less equally, and need to be accounted for to recover the common target belief function $f:\mathcal{X}\to\mathcal{Y}$. Any cost function that results in simplifying the fitted belief function leads to a selection of some dimensions that are paid more attention to, and then, slight differences in the training set lead to a different selections of dimensions.

In this model, it is important that at some high-enough level, the complexity of a belief function makes it prohibitively costly for a person to learn or even express it. This is in line with well-known cognitive limits such as Miller's Law (35), which asserts that the average person only holds about seven features in working memory, as well as a recent study of the interpretability of machine learning models that found evidence that the average person finds it just as difficult to simulate a function of eight or more features as to simulate an opaque "black box" (36). It is natural for agents to form beliefs that achieve a good trade-off between accuracy and simplicity. We show that this trade-off can give rise to significant differences in predictions by learned belief functions even when all of the observed data and the objective reality can be perfectly described by a deterministic function.

We emphasize that settings that lead to disagreement in the complex objective model are not artificial constructs but actually appear in practice. A striking example is the American Housing Survey dataset considered by Mullainathan and Spiess (ref. 37, figure 2). It illustrates this phenomenon precisely in the context of a Lasso selection of variables, which is shown to vary dramatically based on a random sampling of the data.

We also show that in the complex objective model, a distribution that leads to complexity cost-based polarization can be perturbed in some direction such that if agents observe large-enough training sets, then with high probability their learned belief functions will differ only slightly. This suggests that providing some particular information can help refocus agents' selection of variables and lead to consensus. Not any perturbation will do, and the bias needs to be introduced judiciously.

Together, the subjective and objective models provide a foundation for understanding how different sorts of biases in beliefs can arise directly from some constraints or complexity costs in learning.

## Background Definitions

For ease of exposition, for the remainder of the paper we let the set of labels be $\mathcal{Y}=\{-1,+1\}$. The results extend to more general spaces simply by embedding this space in any other that has at least two labels.

**Distributions and Distances.** We consider distributions $\mathcal{D}$ over $\mathcal{X}\times\mathcal{Y}$ and denote by $\mathcal{D}{\downarrow}\mathcal{X}$ the marginal distribution of $\mathcal{D}$ on $\mathcal{X}$. For any two distributions $\mathcal{P}$ and $\mathcal{P}'$ over a domain $\mathcal{X}$, we denote their total variation distance by

$$\mathrm{TV}\left(\mathcal{P},\mathcal{P}'\right):=\sup_{X\subseteq\mathcal{X}}|\mathcal{P}(X)-\mathcal{P}'(X)|.$$

For ease of exposition, we work with the $L_1$ distance between these distributions that is known to satisfy $\|\mathcal{P}-\mathcal{P}'\|_1=2\mathrm{TV}\left(\mathcal{P},\mathcal{P}'\right)$. For any two distributions $\mathcal{D}$ and $\mathcal{D}'$ over $\mathcal{X}\times\mathcal{Y}$, we say that $\mathcal{D}'$ matches the conditional label distributions of $\mathcal{D}$ if for all (but a measure zero under $\mathcal{D}$ and $\mathcal{D}'$) $x\in\mathcal{X}$, $\Pr_{(x,y)\sim\mathcal{D}}[y\mid x]=\Pr_{(x,y)\sim\mathcal{D}'}[y\mid x]$. When $\mathcal{D}'$ matches the conditional label distributions of $\mathcal{D}$, we use $\|\mathcal{D}-\mathcal{D}'\|$ and $\|\mathcal{D}{\downarrow}\mathcal{X}-\mathcal{D}'{\downarrow}\mathcal{X}\|$ interchangeably.

**Belief Functions, Polarization, and Errors.** We consider a belief function class $\mathcal{F}$ of functions $f:\mathcal{X}\to\mathcal{Y}$.

For any belief function $f$, the error of $f$ on $\mathcal{D}$ is described by $\mathrm{err}_{\mathcal{D}}(f):=\Pr_{(x,y)\sim\mathcal{D}}[f(x)\neq y]$.

We say that $\mathcal{D}$ is realizable if there exists $f\in\mathcal{F}$ such that $\mathrm{err}_{\mathcal{D}}(f)=0$.

For a training set $S$ of $m$ labeled input points, $S=\{(x^i,y^i)\}_{i\in[m]}$, we denote the empirical error of $f$ by

$$\mathrm{err}_S(f):=\frac{1}{m}\sum_{i=1}^{m}\mathbb{I}\left(f(x^i)\neq y^i\right).$$

For any $f,f'\in\mathcal{F}$, we denote the disagreement of $f$ and $f'$ on distribution $\mathcal{D}$ by

$$\Delta_{\mathcal{D}}(f,f'):=\Pr_{x\sim\mathcal{D}\downarrow\mathcal{X}}[f(x)\neq f'(x)].$$

For any set of belief functions $\mathcal{H}$, we define the diameter of $\mathcal{H}$, denoted by

$$\mathrm{diam}_{\mathcal{D}}(\mathcal{H}):=\max_{f,f'\in\mathcal{H}}\Delta_{\mathcal{D}}(f,f'),$$

as the largest disagreement between two belief functions in this class. Note that the disagreement between two belief functions and the diameter of a belief function class do not depend on the labels of distribution $\mathcal{D}$. Therefore, with a slight abuse of notation, we use $\mathcal{D}$ in place of $\mathcal{D}{\downarrow}\mathcal{X}$ in these notations or suppress the distribution in the notation for diameter and disagreement when it is clear from context.

We think of a learning setting as being polarizing if agents learn functions whose disagreement is disproportionately larger than the difference between the distributions to which they were exposed. We focus on settings where two agents learn functions $f_1$ and $f_2$ from distributions $\mathcal{D}_1$ and $\mathcal{D}_2$ that have the same marginal distribution $\mathcal{D}$ and either the same or very similar conditional label distribution, yet $\Delta_{\mathcal{D}}(f_1,f_2)$ is large. In this view, polarization is the lack of consensus between agents' beliefs independently of how inaccurate these beliefs may be.

**Sample Complexity.** Let $\mathrm{VCD}(\mathcal{F})$ denote the Vapnik–Chervonenkis dimension (VC dimension) of a belief function class $\mathcal{F}$. That is, $\mathrm{VCD}(\mathcal{F})$ is the cardinality of the largest set of input points $X\subseteq\mathcal{X}$ on which functions in $\mathcal{F}$ can produce all of the $2^{|X|}$ possible labeling. For any $\varepsilon>0$, and $\delta>0$, there is $m_{\varepsilon,\delta}\in O\left(\varepsilon^{-2}\left(\mathrm{VCD}(\mathcal{F})+\ln\left(\frac{1}{\delta}\right)\right)\right)$ such that for any distribution

$\mathcal{D}$ and with probability $1-\delta$ over the choice of set $S$ of at least $m_{\varepsilon,\delta}$ independent and identically distributed (i.i.d.) samples, for all $f \in \mathcal{F}$, we have $|\mathrm{err}_{\mathcal{D}}(f) - \mathrm{err}_S(f)| \le \varepsilon$. When $\mathcal{D}$ satisfies certain properties, one may be able to learn the optimal belief function using fewer samples than presented above. When $\mathcal{D}$ is realizable,

$$m^1_{\varepsilon,\delta} \in O\left(\varepsilon^{-1}\left(\mathrm{VCD}(\mathcal{F})\ln\left(\frac{1}{\varepsilon}\right)+\ln\left(\frac{1}{\delta}\right)\right)\right)$$

is sufficient so that with probability $1-\delta$, any belief function $f$ with empirical error $\mathrm{err}_S(f)=0$ satisfies $\mathrm{err}_{\mathcal{D}}(f) \le \varepsilon$. Even for distributions that are nonrealizable, one can obtain faster learning rates if the Bayes optimal classifier is in $\mathcal{F}$. This property is known as the Massart condition. The statistical and computational aspects of distributions satisfying it have long been of interest (38–41). In particular, if $f^{\mathrm{Bayes}}(x) := \mathrm{argmax}_y \Pr[y|x] \in \mathcal{F}$ and for all $x \in \mathcal{X}$, $|\Pr[y|x] - \Pr[-y|x]| \ge \beta$, then

$$m^{\beta}_{\varepsilon,\delta} \in O\left(\frac{1}{\beta\varepsilon}\left(\mathrm{VCD}(\mathcal{F})\ln\left(\frac{1}{\varepsilon}\right)+\ln\left(\frac{1}{\delta}\right)\right)\right) \qquad \textbf{[1]}$$

samples are sufficient so that with probability $1-\delta$ the empirical error minimizer $\bar{f} \in \mathrm{argmin}_{f\in\mathcal{F}}\mathrm{err}_S(f)$ also satisfies $\mathrm{err}_{\mathcal{D}}(\bar{f}) - \mathrm{err}_{\mathcal{D}}(f^{\mathrm{Bayes}}) \le \varepsilon$.

## The Mixed Subjective Model

In our first model, the mixed subjective model, we represent two world views through two realizable distributions $\mathcal{D}_1$ and $\mathcal{D}_2$ over $\mathcal{X} \times \mathcal{Y}$, which are consistent with belief functions $f_1$ and $f_2$: that is, $\mathrm{err}_{\mathcal{D}_1}(f_1) = \mathrm{err}_{\mathcal{D}_2}(f_2) = 0$. We consider two agents that, perhaps through communication, end up observing training sets from almost identical mixtures of these two distributions and learn belief functions $\widetilde{f}_1$ and $\widetilde{f}_2$; we ask whether these belief functions can be in significant disagreement.

Specifically, consider two agents who start with different distributions $\mathcal{D}_1$ and $\mathcal{D}_2$ that are consistent with two belief functions $f_1$ and $f_2$ with large disagreement $\Delta_{\mathcal{D}}(f_1, f_2)$ (where $\mathcal{D}$ is the shared marginal distribution of $\mathcal{D}_1$ and $\mathcal{D}_2$). Assume that each agent attempts to see the world from the other's perspective; the two agents observe training sets $\widetilde{S}_1$ and $\widetilde{S}_2$ from distributions $\widetilde{\mathcal{D}}_1 := (1-\alpha)\mathcal{D}_1 + \alpha\mathcal{D}_2$ and $\widetilde{\mathcal{D}}_2 := (1-\alpha)\mathcal{D}_2 + \alpha\mathcal{D}_1$, for $\alpha < 1/2$, respectively. Furthermore, assume that these agents choose belief functions $\widetilde{f}_1$ and $\widetilde{f}_2$ that achieve optimal accuracy on training sets $\widetilde{S}_1$ and $\widetilde{S}_2$. Is it possible that $\alpha = 0.49999$ and the two agents, who are learning from almost identical distributions $\widetilde{\mathcal{D}}_1$ and $\widetilde{\mathcal{D}}_2$, would reach very different conclusions?

The following theorem shows that this is not just possible but that they will each stick close to their original belief functions whenever $\alpha < 1/2$: that is, whenever each agent has more weight on their own distribution and the realizable beliefs differ between the two distributions.

**Theorem 1.** *Let distributions $\mathcal{D}_1$ and $\mathcal{D}_2$ on $\mathcal{X} \times \mathcal{Y}$ be two realizable distributions with respect to $\mathcal{F}$ with the same marginal distribution (so we can omit the subscript from $\Delta$ below). Let $\widetilde{\mathcal{D}}_1 := (1-\alpha)\mathcal{D}_1 + \alpha\mathcal{D}_2$ and $\widetilde{\mathcal{D}}_2 := (1-\alpha)\mathcal{D}_2 + \alpha\mathcal{D}_1$, for $\alpha < 1/2$. Then, in the limit as the number of samples grows, the agents' optimal belief functions will converge to the original beliefs and will differ from each other by as much as the originals. That is, for any $\varepsilon > 0$ and $\delta > 0$, there is*

$$m \in O\left(\frac{1}{(\frac{1}{2}-\alpha)^2\varepsilon}\left(\mathrm{VCD}(\mathcal{F})\ln\left(\frac{1}{\varepsilon}\right)+\ln\left(\frac{1}{\delta}\right)\right)\right)$$

*such that if sample sets $S_1$, $S_2$, $\widetilde{S}_1$, and $\widetilde{S}_2$ of size at least $m$ are sampled from $\mathcal{D}_1$, $\mathcal{D}_2$, $\widetilde{\mathcal{D}}_1$, and $\widetilde{\mathcal{D}}_2$, respectively, then with probability at least $1-\delta$, $\Delta(\widetilde{f}_1, \widetilde{f}_2) \ge \Delta(f_1, f_2) - \varepsilon$, $\Delta(\widetilde{f}_1, f_1) \le \varepsilon/4$, and $\Delta(\widetilde{f}_2, f_2) \le \varepsilon/4$, where $f_i \in \mathrm{argmin}_{f\in\mathcal{F}}\mathrm{err}_{S_i}(f)$ and $\widetilde{f}_i \in \mathrm{argmin}_{f\in\mathcal{F}}\mathrm{err}_{\widetilde{S}_i}(f)$ for $i \in \{1,2\}$.*

The proof of *Theorem 1* is given in *SI Appendix*, section A. At a high level, the reason for this phenomenon is that belief functions $f_1$ and $f_2$ remain the optimal belief functions for distributions $\widetilde{\mathcal{D}}_1$ and $\widetilde{\mathcal{D}}_2$. This is due to the fact that for any $(x,y) \sim \mathcal{D}_i$, $(x,y)$ appears with higher probability than $(x,-y)$ in $\widetilde{\mathcal{D}}_i$. So, the optimal classifier for $\widetilde{\mathcal{D}}$ should label $x$ just as the perfect belief function would label it on $\mathcal{D}_i$. The dependence of the sample size $m$ on $\varepsilon$ and $\alpha$ shows that the more instances the agents observe, the more certain they become of their original belief functions. The sample complexity bounds used in this theorem are nearly tight. This means that the more observations agents have, the more likely it is for them to polarize since their samples match $\widetilde{\mathcal{D}}_1$ and $\widetilde{\mathcal{D}}_2$ more, respectively. Perhaps paradoxically, in this model as people gain more information—for instance, with technological advances—they become more likely to polarize.

We note that $f_1$ and $f_2$ could also be found as being the most likely functions to explain data from distributions $\widetilde{\mathcal{D}}_1$ and $\widetilde{\mathcal{D}}_2$, respectively, if one started with a (uniform) prior over belief functions. More generally, any "consistent" learning method—that chooses functions with as little error as possible—leads to the conclusion of *Theorem 1*.

The theorem applies to situations in which the two agents see different labeling for the same inputs. Even if the distributions they see are nearly the same, slight differences in the frequencies of labels that they end up observing for the same inputs are enough to allow them to reach different conclusions in terms of the error-minimizing belief functions that they adopt. It is in this sense that even people who strive to communicate and find common ground with others can form polarized opinions. This gives one possible explanation for polarization in beliefs, predicated upon differences in experiences.

An important aspect of *Theorem 1* is that the distributions that the agents see, $\mathcal{D}_1$ and $\mathcal{D}_2$, agree. This effectively means that they are examining the same issues. If these were instead disjoint—for instance, with one agent being an expert on a topic related to chemistry and the other on a topic related to linguistics, with no knowledge of the other topic—then when they communicated they would then simply adopt the other's belief function on the other's topic. The interest in *Theorem 1* comes from the fact that the agents are exchanging information about a topic on which they both have previous information and beliefs.

We next explore a different explanation for how agents can arrive at very different belief functions despite a large overlap in information.

## The Complex Objective Model

In this section, we turn our attention to the complex objective model, where the common world view is represented by a realizable distribution $\mathcal{D}$ that is consistent with some belief function $f^* \in \mathcal{F}$. In this case, any agent who observes a sufficiently large number of labeled samples from $\mathcal{D}$ and adopts the belief function that minimizes error on these samples learns a belief that is in almost full agreement with $f^*$. Therefore, all error-minimizing agents will arrive at belief functions that are almost in full agreement with each other.

However, when the error-minimizing belief functions are very complex, the agent may face difficulty learning or interpreting them. Therefore, rather than considering error-minimizing agents, we consider agents who attempt to find a belief function that strikes a balance between accuracy and complexity. We first show that when there is a complexity associated with learning functions, it is entirely possible that two agents receiving two i.i.d.

training sets from $\mathcal{D}$ would learn belief functions $f_1$ and $f_2$ that are in large disagreement. We then discuss how this can be dealt with by changing the original distribution $\mathcal{D}$ slightly to help the two agents arrive at belief functions that are mostly in agreement.

**Setup and Initial Observations.** Before we proceed, we require some notation.

$\mathcal{F}$ is accompanied by a complexity cost $\phi(\cdot)$, such that for any $f \in \mathcal{F}$, $\phi(f)$ determines how complex the belief function $f$ is and how much that costs the decision maker. Some examples of such complexity costs include monotonic functions of the number of features in a Boolean function or the depth of a decision list.

For a distribution $\mathcal{D}$ and a training set $S$, we denote the overall cost incurred by a belief function $f \in \mathcal{F}$ through its error and complexity cost as

$$\text{cost}_{\mathcal{D}}(f) := \text{err}_{\mathcal{D}}(f) + \phi(f) \ \text{ and } \text{cost}_S(f) := \text{err}_S(f) + \phi(f).$$

We first demonstrate the phenomenon described above by providing a simple example in which there are at least two optimal belief functions $f_1, f_2 \in \operatorname{argmin}_{f\in\mathcal{F}} \text{cost}_{\mathcal{D}}(f)$ that have a nontrivial disagreement with each other.

***Example 1*:** Let $\mathcal{X} := \mathbb{R}^d$, and let $\mathcal{F}$ be the class of homogeneous linear separators: that is, $\mathcal{F} := \{f_{\mathbf{w}}(\mathbf{x}) = \text{sign}(\mathbf{w}\cdot\mathbf{x})\}_{\mathbf{w}\in\mathbb{R}^d}$, where $\text{sign}(z) = +1$ for $z \geq 0$ and $-1$ otherwise. Let $\mathcal{D}$ be the simple distribution that labels positive unit vectors by $+1$ and negative unit vectors by $-1$, with equal weight on each. That is, $\mathcal{D}$ has weight $\frac{1}{2d}$ on each of $(\mathbf{e}_i, +1)$ and $(-\mathbf{e}_i, -1)$ for all $i \in \{1, \ldots, d\}$, where $\mathbf{e}_i$ is the $i$th unit vector.

Let the complexity cost be $\phi(f_{\mathbf{w}}) := h\left(\|\mathbf{w}\|_0\right)$ for $h$ such that $h(0) = 0$, $h(1) \leq \frac{1}{2d}$, and $h(d) \geq \frac{1}{2}$. It then follows that there exists at least one $k \in \{1, \ldots, d-1\}$ for which all of the functions with $k$ dimensions are optimal:

$$\left\{\mathbf{w} \in \{0,1\}^d \mid \|\mathbf{w}\|_0 = k\right\} \subseteq \operatorname*{argmin}_{\mathbf{w}\in\mathbb{R}^d} \text{cost}_{\mathcal{D}}(f_{\mathbf{w}}). \qquad \textbf{[2]}$$

To see that Eq. **2** holds, first note that for any $\mathbf{w} \in \mathbb{R}^d$, rounding $w_i > 0$ to 1 and $w_i < 0$ to $-1$ does not change its error or cost. Furthermore, for any $i$, if $w_i = 1$, then $f_{\mathbf{w}}$ labels both $(\mathbf{e}_i, +1)$ and $(-\mathbf{e}_i, -1)$ correctly; if $w_i = 0$, then $f_{\mathbf{w}}$ labels $(\mathbf{e}_i, +1)$ correctly and $(-\mathbf{e}_i, -1)$ incorrectly, and if $w_i = -1$, then $f_{\mathbf{w}}$ labels both $(\mathbf{e}_i, +1)$ and $(-\mathbf{e}_i, -1)$ incorrectly. Thus, we can restrict our attention to $\mathbf{w} \in \{0,1\}^d$. Given that setting $w_i = 1$ instead of $w_i = 0$ decreases the error by exactly $\frac{1}{2d}$, it follows that if $\|\mathbf{w}\|_0 = \|\mathbf{w}'\|_0$, then $\phi(f_{\mathbf{w}}) = \phi(f_{\mathbf{w}'})$. The facts that $h(0) = 0$, $h(1) \leq \frac{1}{2d}$, and $h(d) \geq \frac{1}{2}$ imply that the overall cost of $\mathbf{w}$ with a single nonzero feature is at most that of setting $\mathbf{w}$ to be the vector of all zeros or all ones. Therefore, [**2**] holds for at least one $k \in \{1, \ldots, d-1\}$.

Next, note that any two belief functions that separate on different dimensions and both have the optimal overall number of dimensions $k$ must have a minimal distance between them. Specifically, if $\mathbf{w}, \mathbf{w}' \in \{0,1\}^d$ are such that $\|\mathbf{w}\|_0 = \|\mathbf{w}'\|_0 = k$ and $\mathbf{w} \neq \mathbf{w}'$, then $\frac{1}{d} \leq \Delta_{\mathcal{D}}(f_{\mathbf{w}}, f_{\mathbf{w}'})$.

We emphasize that although the example is special in the symmetry of $\mathcal{D}$, that is not necessary. One could have very different gains from classifying correctly on some dimensions. What is necessary is that there is more than one choice for feature subsets that have the same complexity cost and the same predictive power. This can happen when the marginal predictive power of including an additional feature is the same for multiple features. A good example of this situation, where the marginal benefit of different features in a classifier is nearly the same, is seen in practice by Mullainathan and Spiess (37).

*Example 1* implies that there are situations in which it is likely that two agents would learn belief functions that disagree substantially, as shown in the following theorem.

**Theorem 2.** *There is a distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$ that is realizable with respect to a belief function class $\mathcal{F}$ and a complexity cost function $h$, such that for any $m$ and two sets of $m$ i.i.d. samples, $S_1$ and $S_2$, with probability at least $\frac{1}{4}$, there are $f_i \in \operatorname{argmin}_{f\in\mathcal{F}} \text{cost}_{S_i}(f)$ for $i \in \{1,2\}$ such that $\Delta_{\mathcal{D}}(f_1, f_2) > \frac{1}{6}$.*

***Proof of Theorem 2*:** We use the setting of *Example 1* and let $h(k) = \frac{k}{2d}$. Note that if $S_1$ has at most a $\frac{1}{2d}$ fraction of its samples on $(-\mathbf{e}_j, -1)$, then there is a cost-minimizing $f_1$ that has $w_j = 0$ and $w_j = 1$ otherwise. Similarly, if $S_2$ has at least a $\frac{1}{2d}$ fraction of its samples on $(-\mathbf{e}_j, -1)$, then there is a cost-minimizing $f_2$ that has $w_j = 1$ and $w_j = 0$ otherwise. Using well-known properties of binomial distributions (42), it follows that the probability that $f_1(-\mathbf{e}_j) \neq f_2(-\mathbf{e}_j)$ is at least $1/2$. [Note that using specific $f_i \in \operatorname{argmin}_{f\in\mathcal{F}} \text{cost}_{S_i}(f)$ allows us to break ties in a way that is convenient for us. This is purely for ease of exposition; we could formulate the theorem for any cost-minimizing $f_i$, using slightly messier probability calculations.]

Let $Z$ be the random variable whose value is the number of coordinates $j \in [d]$ such that $f_1(-\mathbf{e}_j) \neq f_2(-\mathbf{e}_j)$. It holds that $Z \leq d$, and $\mathbb{E}[Z] \geq d/2$. Therefore, it must be the case that $\Pr[Z \leq d/3] \leq 3/4$: that is, $\Pr[Z > d/3] > 1/4$. Hence, with probability at least $1/4$, $\Delta_{\mathcal{D}}(f_1, f_2) > 1/6$. □

*Proof of Theorem 2* uses a specific linear complexity cost that makes the probability statement clean and the proof straightforward. However, it is extends to much more general cost functions. For example, instead of having a linear cost function where the cost of each additional dimension exactly balances the expected benefit of separation, we can work with any complexity cost function $h$ that makes it optimal to choose some dimension $k \in \{1, \ldots, d-1\}$, such as in *Example 1*. In that more general case, there is still a probability bounded away from zero that the two optimal functions would be different. Indeed, in *SI Appendix*, section D, Theorem 5, we show that $\Delta_{\mathcal{D}}(f_1, f_2) \geq \frac{1}{2d}$ with probability at least $5/12$ (for any training set size and $d \geq 3$; we also show that the probability goes to 1 as $d$ increases).

The setting of *Example 1* and *Proof of Theorem 2* is constructed to most directly demonstrate how accounting for costs provides an explanation for polarization. The intuition behind the example and theorem makes it clear that polarization is something that can easily arise after the setting involves many dimensions and complexity leads people to choose some subset. The chance that they coordinate on exactly the same dimensions is only high if those dimensions do much better than other dimensions. Whenever there are multiple dimensions that have similar importance in determining outcomes, different observers can easily favor different dimensions in their belief functions. Again, we note that nontrivial levels of disagreement between learned models have been observed in practice when there are many potential variables to include in finding a good classifier.

A possible explanation for why polarization occurs even under "natural" distributions is that perturbing the polarizing distributions by some unbiased noise does not alleviate the problem (e.g., ref. 43). That is, adding unbiased (but realizable) noise to the above example, such as adding a uniform distribution over the domain that is labeled by the optimal belief function, still leads to the same outcome. This is best observed by noting that distribution $\mathcal{D}$ of *Example 1* is uniform over its support, so noise that does not introduce bias toward specific subsets of the support does not change the distribution at all. Our next result builds on this intuition.

**Introducing Bias to Prevent Polarization.** We argue that for any problematic distribution $\mathcal{D}$ as in *Theorem 2*, one can carefully design $\widetilde{\mathcal{D}}$ that is close to $\mathcal{D}$ in its marginal distribution and is still realizable with respect to $\mathcal{F}$, such that $\widetilde{\mathcal{D}}$ does not suffer from the same problem. That is, it is always possible to prevent polarization (of the type studied here) by introducing a slight bias into the information selection process. Formally, we prove the following theorem.

**Theorem 3.** *Consider a function class $\mathcal{F}$, a distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$ labeled by an $f^* \in \mathcal{F}$, an $\alpha \in (0,1]$, and a maximum level of disagreement $\gamma > 0$. Then, there is a distribution $\widetilde{\mathcal{D}}$ labeled by $f^*$ such that $\|\widetilde{\mathcal{D}} - \mathcal{D}\|_1 \leq \alpha$, and there is*

$$m \in O\left(\gamma^{-4}\alpha^{-2}\left(\mathrm{VCD}(\mathcal{F}) + \ln\left(\frac{1}{\delta}\right)\right)\right),$$

*such that if two sets $S_1$ and $S_2$ of size at least $m$ are sampled from $\widetilde{\mathcal{D}}$, then with probability at least $1-\delta$, any two cost-minimizing belief functions $f_i \in \mathrm{argmin}_{f\in\mathcal{F}} \mathrm{cost}_{S_i}(f)$ for $i \in \{1,2\}$:*

1) *have at most $\gamma$ disagreement over $\mathcal{D}$ [i.e., $\Delta_{\mathcal{D}}(\widetilde{f}_1, \widetilde{f}_2) \leq \gamma$] and*
2) *have a cost that is optimal up to $3\alpha$ on $\mathcal{D}$: that is, $\mathrm{cost}_{\mathcal{D}}(\widetilde{f}_i) \leq \mathrm{argmin}_{f\in\mathcal{F}} \mathrm{cost}_{\mathcal{D}}(f) + 3\alpha$.*

*Theorem 3* states that even if $\mathcal{D}$ is a polarizing distribution, there is a nearby distribution $\widetilde{\mathcal{D}}$ that is not polarizing. We can think of $\widetilde{\mathcal{D}}$ as an intervention. As an example, $\mathcal{D}$ could correspond the choices of news articles curated by a news agency or the social media posts that appear on a person's news feed. An intervention in this case refers to a (small) increase in the frequency of some types of content, which leads to a less polarizing distribution $\widetilde{\mathcal{D}}$ that is still close to $\mathcal{D}$. As mentioned earlier, this intervention has to be carefully chosen so that it removes the symmetries in $\mathcal{D}$ that can cause samples from the distribution to differ in ways that lead to significantly differently belief functions. After such symmetries are eliminated, large samples are likely to lead to the same belief function.

The proof of *Theorem 3* is relegated to *SI Appendix*, section A. Here, we provide an overview of this proof.

***Sketch of the proof of Theorem 3.*** First, to assist with examining disagreement between cost-minimizing belief functions, we introduce some additional notation. For any $\varepsilon$ and $\mathcal{D}$, let

$$\mathcal{F}_\varepsilon^{\mathcal{D}} := \left\{ f \in \mathcal{F} \;\middle|\; \mathrm{cost}_{\mathcal{D}}(f) \leq \underset{f'\in\mathcal{F}}{\mathrm{argmin}}\, \mathrm{cost}_{\mathcal{D}}(f') + \varepsilon \right\}.$$

This definition is loosely related to the concept of Rashomon sets from statistics (44).

If $\mathrm{diam}(\mathcal{F}_\varepsilon^{\mathcal{D}}) \leq \gamma$ is small, then polarization is not an issue (i.e., two agents with sufficiently many samples from $\mathcal{D}$ will learn belief functions with disagreement of at most $\gamma$). This follows from the next lemma (with $\mathcal{D} = \mathcal{D}'$), whose proof appears in *SI Appendix*, section B.

**Lemma 1.** *Consider two distributions $\mathcal{D}$ and $\mathcal{D}'$, such that $\mathrm{diam}_{\mathcal{D}}\left(\mathcal{F}_\varepsilon^{\mathcal{D}'}\right) \leq \gamma$. Then, there is*

$$m \in O\left(\varepsilon^{-2}\left(\mathrm{VCD}(\mathcal{F}) + \ln\left(\frac{1}{\delta}\right)\right)\right)$$

*such that for two sample sets $S_1$ and $S_2$ of size at least $m$ from distribution $\mathcal{D}'$, with probability $1-\delta$, if $f_i \in \mathrm{argmin}_{f\in\mathcal{F}} \mathrm{cost}_{S_i}(f)$ for $i \in \{1,2\}$, it holds that $f_1, f_2 \in \mathcal{F}_\varepsilon^{\mathcal{D}'}$ and $\Delta_{\mathcal{D}}(f_1, f_2) \leq \gamma$.*

Unfortunately, $\mathrm{diam}_{\mathcal{D}}(\mathcal{F}_\varepsilon^{\mathcal{D}})$ can be large even for small values of $\varepsilon$—see *Example 1*, where $\mathrm{diam}_{\mathcal{D}}(\mathcal{F}_0^{\mathcal{D}}) = \frac{1}{2}$. This makes polarization unavoidable when two agents learn from distribution $\mathcal{D}$.

However, we show that for any $\mathcal{D}$ such that $\mathrm{diam}_{\mathcal{D}}(\mathcal{F}_\varepsilon^{\mathcal{D}})$ is large, there is a distribution $\widetilde{\mathcal{D}}$ close to $\mathcal{D}$ with a much smaller $\mathrm{diam}_{\mathcal{D}}(\mathcal{F}_\varepsilon^{\widetilde{\mathcal{D}}})$. This implies that learning over the distribution $\widetilde{\mathcal{D}}$ aligns the learning process of the two agents and leads to models that have small disagreement. More details are in the next lemma, whose proof appears in *SI Appendix*, section B.

**Lemma 2.** *Let $\mathcal{D}$ be a distribution over $\mathcal{X} \times \mathcal{Y}$ that is realizable with respect to $\mathcal{F}$ [i.e., there is $f^* \in \mathcal{F}$ such that $\mathrm{err}_{\mathcal{D}}(f^*) = 0$]. Assume that for some $\varepsilon > 0$, $\mathrm{diam}_{\mathcal{D}}\left(\mathcal{F}_\varepsilon^{\mathcal{D}}\right) \geq 2R$ for some $R \in [0, 0.5]$. Then, there is a distribution $\mathcal{P}$ with $\mathrm{err}_{\mathcal{P}}(f^*) = 0$, such that for any $\alpha > 0$ and $\widetilde{\mathcal{D}} = (1-\alpha)\mathcal{D} + \alpha\mathcal{P}$, $\mathrm{diam}_{\mathcal{D}}\left(\mathcal{F}_\varepsilon^{\widetilde{\mathcal{D}}}\right) \leq \frac{8\varepsilon}{\alpha R}$.*

At a high level, the proof of *Lemma 2* follows by designing a distribution $\widetilde{\mathcal{D}}$ that biases $\mathcal{D}$ toward a specific belief function $\widetilde{f}$. An interesting aspect of this $\widetilde{f}$ is that it is one of the least complex among the belief functions that have near-optimal cost, $\mathcal{F}_\varepsilon^{\mathcal{D}}$. This biases the agents' learning process toward less complex but also, less accurate models. In *SI Appendix*, section B.1, we provide more insights on the choice of $\widetilde{f}$ and show that such a trade-off between simplicity and accuracy is unavoidable when we aim to reduce polarization.

Lastly, the proof of *Theorem 3* follows from the above lemmas. At a high level, if $\mathrm{diam}_{\mathcal{D}}(\mathcal{F}_\varepsilon^{\mathcal{D}})$ is sufficiently small, then by *Lemma 1*, $\mathcal{D}$ is not a polarizing distribution. On the other hand, if $\mathrm{diam}_{\mathcal{D}}(\mathcal{F}_\varepsilon^{\mathcal{D}})$ is large, then by *Lemma 2* there is $\widetilde{\mathcal{D}}$ close to $\mathcal{D}$ such that $\mathrm{diam}_{\mathcal{D}}(\mathcal{F}_\varepsilon^{\widetilde{\mathcal{D}}})$ is small. Using *Lemma 1*, two agents who sample from $\widetilde{\mathcal{D}}$ learn belief functions that are in close agreement on $\mathcal{D}$. Moreover, $\widetilde{\mathcal{D}}$ is close to $\mathcal{D}$, so these belief functions are almost optimal with respect to $\mathcal{D}$.

**Lower Bound.** As we observed in *Theorem 3*, for any desired maximum level of disagreement between two agents, $\gamma$, and any desired level of intervention, $\alpha$, every distribution $\mathcal{D}$ can be changed to a nearby distribution $\mathcal{D}'$ at distance $\alpha$, such that agents who receive a large-enough number of samples from $\mathcal{D}'$ have disagreement of less than $\gamma$. *Theorem 3* shows that the number of samples needed for this to work is at most $O\left(\gamma^{-4}\alpha^{-2}\mathrm{VCD}(\mathcal{F})\right)$. We next provide a lower bound that shows that the number of samples needed for the agents to avoid polarization indeed has to increase with $\frac{1}{\alpha}$ and $\frac{1}{\gamma}$. That is, we succeed at having smaller disagreement between agents and making only small change to the distribution only if agents form their belief functions after having acquired a large number of observations.

**Theorem 4.** *Let $m(\alpha, \gamma, d, \delta)$ be as follows. For any distribution $\mathcal{D}$ on domain $\mathcal{X}$ and any belief function class $\mathcal{F} \subseteq \mathcal{Y}^{\mathcal{X}}$ with VC dimension $d$ and any cost function $\phi$ over $\mathcal{F}$, there exists $\mathcal{D}'$ such that $\|\mathcal{D} - \mathcal{D}'\|_1 \leq \alpha$, $\mathcal{D}$ and $\mathcal{D}'$ have the same conditional label distributions, and for any $m \geq m(\alpha, \gamma, d, \delta)$, with probability $1-\delta$, $\Delta_{\mathcal{D}}(f_1, d_2) \leq \gamma$, where $f_i \in \mathrm{argmin}_{f\in\mathcal{F}} \mathrm{cost}_{S_i}(f)$ and $S_1$ and $S_2$ are two i.i.d. sample sets of size $m$ from $\mathcal{D}'$. Then, for any $\alpha < \frac{1}{3}$, $\gamma < \frac{1}{2}$, and $d \geq 1/\gamma$, we have that*

$$m\left(\alpha, \gamma, d, \frac{1}{4}\right) \in \Omega\left(\frac{d}{\alpha^2}\ln\left(\frac{1}{\gamma}\right)\right).$$

The proof of *Theorem 4* appears in SI Appendix, section C. We remark that there is a gap in terms of the dependence on parameter $\frac{1}{\gamma}$ between the upper bound (*Theorem 3*) and the lower bound (*Theorem 4*). This is perhaps good news. After all, we may be able to avoid polarization with a significantly smaller number of samples than that prescribed by *Theorem 3*. In *SI Appendix*, section D, we discuss in more detail the source of this gap between the upper and lower bounds and present a possible path forward toward an improved upper bound.

## Discussion

Our results show that polarization that arises from differences in subjective opinion is unlike polarization that arises from difficulty in processing objective information. Indeed, the mixed subjective model is pessimistic in that polarization can be inevitable. By contrast, in the complex objective model, our result (*Theorem 3*) is more positive; even though polarization arises, we can always introduce a slight bias into the information selection process (i.e., perturb the distribution) in a way that leads to consensus. While our model is admittedly a stylized abstraction of reality, the conceptual message is appealing; in some situations, small interventions can eliminate polarization.

More generally, the world may have aspects of both the subjective and objective models. People may develop models over some complex intersection that are also optimal in describing past experiences that are more idiosyncratic. The dimensions that they focus on when faced with some overlapping common and objective data may have been heavily influenced by their own personal experiences outside of that domain. This suggests that to get people to incorporate common dimensions may require providing them with data that explicitly shows them the value of incorporating those dimensions into their beliefs. Convincing someone that the climate is changing may be more reliably accomplished by showing them the value of science in some other domain and getting them to trust science as an explanatory factor, rather than showing them more data on climate.

We note that although our agents polarize when considering similar data, both subjectively or objectively, they end up with functions that are equally optimal. Thus, the reason for wanting consensus in our settings does not come from correcting some agents who are suboptimal. The motivation for consensus comes from the fact that polarization in beliefs leads agents to prefer different actions or policies, which can give rise to disagreement and even conflict in politics and collective decision making. There is evidence that various divisions and forms of polarization lead to lower growth and other forms of inefficiencies (45–47). Therefore, reaching a consensus can be valuable in and of itself.

One may wonder whether Bayesian models can extend to our setting. Indeed, some of the results from our mixed subjective model remain the same if the learner starts from a uniform prior on $\mathcal{F}$ and picks the most likely function to explain the observed data. However, this approach does not naturally accommodate the complexity cost of learning functions. In addition, our approach differs from a Bayesian approach on at least two conceptual levels. First, a Bayesian model begins with a full specification of the world, refines it, and cannot make predictions outside of its conceived possibilities. By contrast, an aspect of a learning-theoretic model for human learning is that it allows people to "generalize"—that is, form beliefs for new situations that they encounter that go beyond their previous conception. In other words, our model naturally allows for a person to learn across different experiences and to form opinions $f(x)$ about unseen instances $x$ solely based on prior experience; essentially reasoning by analogy from their experiences, the function can make predictions for specific instances that they have not experienced or conceived of based on reasoning from instances that they have seen. For instance, if $\mathcal{F}$ is a class of linear functions and one had not previously conceived that certain combinations of $x$ values are possible, one can still choose a new function from that class that best fits the expanded observations. If those are outside of the prior distribution and one is Bayesian, then there is no prediction. Second, on a more descriptive side, as we discussed earlier, in our approach if a person is faced with new data, they can change their world view to extend their class of belief functions (for instance, linear ones) to fit over a larger domain and adopt a completely new belief function $f$ that best matches the new data. Thus, learning can involve completely changing one's model of the world, and there is ample evidence from developmental psychology, for instance, that children develop a new understanding of the world as they grow. By contrast, Bayesian modeling would posit that people are born with prior distributions over all possible models of the world that they refine throughout their lifetimes.

Our results have limitations. Our negative result for the mixed subjective model (*Theorem 1*) relies on the assumption that agents choose a deterministic belief function, rather than a probabilistic belief. Although it is clear that a qualitatively similar result would hold when agents' beliefs are "close" to being deterministic, it is unclear how the result generalizes to probabilistic beliefs. In addition, recall that our positive result for the complex objective model (*Theorem 3*) provides an intervention that is valid even in the worst case. A shortcoming of *Theorem 3*, especially in terms of making this message more practical, is that the intervention needs to be tailored to the setting. It requires knowledge of the belief function class, the cost function, and the composition of the set $\mathcal{F}_{\varepsilon}^{\mathcal{D}}$. An important open question, therefore, is whether it is possible to design this intervention through a simple and tractable algorithm that does not explicitly construct $\mathcal{F}_{\varepsilon}^{\mathcal{D}}$.

As a closing remark, in recent years social scientists have started to embrace machine learning. The theory developed here shows that machine learning not only provides a set of methods for empirical work but can also provide a foundation for modeling human belief formation and decision making and developing insights into things like polarization.

**Data Availability.** There are no data underlying this work.

1. D. G. McNeil Jr., Measles cases surpass 700 as outbreak continues unabated. *New York Times*, 29 April 2019. https://nyti.ms/2L8mzmP. Accessed 1 May 2020.
2. C. McCoy, Anti-vaccination beliefs don't follow the usual political polarization. *Conversation*, 23 August 2017. https://goo.gl/4gMCQP. Accessed 1 May 2020.
3. M. McPherson, L. Smith-Lovin, J. M. Cook, Birds of a feather: Homophily in social networks. *Annu. Rev. Sociol.* **27**, 415–444 (2001).
4. B. Golub, M. O. Jackson, How homophily affects the speed of learning and best-response dynamics. *Q. J. Econ.* **127**, 1287–1338 (2012).
5. E. Bakshy, S. Messing, L. A. Adamic, Exposure to ideologically diverse news and opinion on Facebook. *Science* **348**, 1130–1132 (2015).
6. C. R. Sunstein, *#Republic: Divided Democracy in the Age of Social Media* (Princeton University Press, 2018).
7. M. O. Jackson, *The Human Network: How Your Social Position Determines Your Power, Beliefs, and Behaviors* (Pantheon Books, New York, 2019).
8. C. G. Lord, L. Ross, M. R. Lepper, Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *J. Pers. Soc. Psychol.* **37**, 2098 (1979).
9. R. G. Fryer, P. Harms, M. O. Jackson, Updating beliefs when evidence is open to interpretation: Implications for bias and polarization. *J. Eur. Econ. Assoc.* **17**, 1470–1501 (2019).
10. C. A. Bail *et al.*, Exposure to opposing views on social media can increase political polarization. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 9216–9221 (2018).
11. A. Alesina, A. Miano, S. Stantcheva, "The polarization of reality" in *AEA Papers and Proceedings*, W. R. Johnson, G. Herbert, Eds. (AEA, 2020), vol. 110, pp. 324–328.
12. R. S. Nickerson, Confirmation bias: A ubiquitous phenomenon in many guises. *Rev. Gen. Psychol.* **2**, 175–220 (1998).
13. M. Rabin, J. L. Schrag, First impressions matter: A model of confirmatory bias. *Q. J. Econ.* **114**, 37–82 (1999).
14. J. von Neumann, O. Morgenstern, *Theory of Games and Economic Behavior* (Princeton University Press, 1944).
15. L. J. Savage, *The Foundations of Statistics* (Wiley, Oxford, United Kingdom, 1954).
16. J.-P. Benoit, J. Dubra, A theory of rational attitude polarization. SSRN [Preprint] (2014). dx.doi.org/10.2139/ssrn.2529494 (Accessed 1 May 2020).
17. A. G. Chandrasekhar, H. Larreguy, J. P. Xandri, Testing models of social learning on networks: Evidence from a lab experiment in the field. *Econometrica* **88**, 1–32 (2020).
18. R. R. Bush, F. Mosteller, *Stochastic Models for Learning* (John Wiley & Sons, 1955).
19. M. H. DeGroot, Reaching a consensus. *J. Am. Stat. Assoc.* **69**, 118–121 (1974).
20. B. L. Lipman, Information processing and bounded rationality: A survey. *Can. J. Econ.* **28**, 42–67 (1995).
21. M. Mobius, T. Rosenblat, Social learning in economics. *Annu. Rev. Econ.* **6**, 827–847 (2014).
22. B. Golub, E. Sadler, "Learning in social networks" in *The Oxford Handbook of the Economics of Networks*, Y. Bramoullé, A. Galeotti, B. Rogers, Eds. (Oxford University Press, 2016), pp. 504–542.
23. R. S. Sutton, A. G. Barto, *Reinforcement Learning: An Introduction* (MIT Press, 2018).

24. J. A. Bohren, D. N. Hauser, Social learning with model misspecification: A framework and a characterization. SSRN [Preprint] (2019). dx.doi.org/10.2139/ssrn.3236842 (Accessed 1 May 2020).
25. R. Spiegler, Behavioral implications of causal misperceptions. *Annu. Rev. Econ.* **12**, 80–106 (2020).
26. L. Smith, P. Sørensen, Pathological outcomes of observational learning. *Econometrica* **68**, 371–398 (2000).
27. E. Mossel, M. Mueller-Frank, A. Sly, O. Tamuz, Social learning equilibria. *Econometrica* **88**, 1235–1267 (2020).
28. P. Rochat, Five levels of self-awareness as they unfold early in life. *Conscious. Cognit.* **12**, 717–731 (2003).
29. H. A. Simon, *Administrative Behavior: A Study of Decision-Making Processes in Administrative Organization* (Macmillan, 1947).
30. C. H. Papadimitriou, M. Yannakakis, "On complexity as bounded rationality" in *Proceedings of the Twenty-Sixth Annual ACM Symposium on Theory of Computing*, F. T. Leighton, M. T. Goodrich, Eds. (ACM, 1994), pp. 726–733.
31. A. Rubinstein, *Modeling Bounded Rationality* (MIT Press, 1998).
32. R. Radner, "Can bounded rationality resolve the prisoner's dilemma" in *Essays in Honor of Gerard Debreu*, A. Mas-Colell, W. Hildenbrand, Eds. (North-Holland, Amsterdam, The Netherlands, 1986), pp. 387–399.
33. E. Kalai, "Bounded rationality and strategic complexity in repeated games" in *Game Theory and Applications*, T. Ichiishi, A. Neyman, Y. Tauman, Eds. (Elsevier, 1990), pp. 131–157.
34. L. G. Valiant, A theory of the learnable. *Commun. ACM* **27**, 1134–1142 (1984).
35. G. A. Miller, The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychol. Rev.* **101**, 343–352 (1956).
36. F. Poursabzi-Sangdeh, D. G. Goldstein, J. M. Hofman, J. W. Vaughan, H. M. Wallach, Manipulating and measuring model interpretability. arXiv [Preprint] (2018). https://arxiv.org/abs/1802.07810 (Accessed 1 May 2020).
37. S. Mullainathan, J. Spiess, Machine learning: An applied econometric approach. *J. Econ. Perspect.* **31**, 87–106 (2017).
38. O. Bousquet, S. Boucheron, G. Lugosi, Theory of classification: A survey of recent advances. *ESAIM Probab. Stat.* **9**, 323–375 (2005).
39. P. Awasthi, M. Balcan, N. Haghtalab, R. Urner, "Efficient learning of linear separators under bounded noise" in *Proceedings of the 28th Conference on Computational Learning Theory (COLT)*, P. Grünwald, E. Hazan, S. Kale, Eds. (JMLR, 2015), pp. 167–190.
40. P. Awasthi, M. F. Balcan, N. Haghtalab, H. Zhang, "Learning and 1-bit compressed sensing under asymmetric noise" in *Proceedings of the 29th Conference on Computational Learning Theory (COLT)*, V. Feldman, A. Rakhlin, O. Shamir, Eds. (JMLR, 2016), pp. 152–192.
41. E. Mammen, A. B. Tsybakov, Smooth discrimination analysis. *Ann. Stat.* **27**, 1808–1829 (1999).
42. K. Jogdeo, S. Samuels, Monotone convergence of binomial probabilities and a generalization of Ramanujan's equation. *Ann. Math. Stat.* **39**, 1191–1195 (1968).
43. D. A. Spielman, S. H. Teng, Smoothed analysis of algorithms: Why the simplex algorithm usually takes polynomial time. *J. ACM* **51**, 385–463 (2004).
44. A. Fisher, C. Rudin, F. Dominici, All models are wrong, but many are useful: Variable importance for black-box, proprietary, or misspecified prediction models, using model class reliance. arXiv [Preprint] (2018). https://arxiv.org/abs/1801.01489v3 (Accessed 1 May 2020).
45. P. Keefer, S. Knack, Polarization, politics and property rights: Links between inequality and growth. *Publ. Choice* **111**, 127–154 (2002).
46. A. Alesina, A. Devleeschauwer, W. Easterly, S. Kurlat, R. Wacziarg, Fractionalization. *J. Econ. Growth* **8**, 155–194 (2003).
47. A. Alesina, E. L. Ferrara, Ethnic diversity and economic performance. *J. Econ. Lit.* **43**, 762–800 (2005).