When learning becomes impossible

Nicholas Asher
asher@irit.fr
CNRS, IRIT
Toulouse, France

Julie Hunter
jhunter@linagora.com
LINAGORA Labs
Toulouse, France

ABSTRACT

We formally analyze an epistemic bias we call *interpretive blindness* (IB), in which under certain conditions a learner will be incapable of learning. IB is now common in our society, but it is a natural consequence of Bayesian inference and what we argue are mild assumptions about the relation between belief and evidence. IB a special problem for learning from testimony, in which one acquires information only from text or conversation. We show that IB follows from a codependence between background beliefs and interpretation in a Bayesian setting and the nature of contemporary testimony. We argue that a particular characteristic of contemporary testimony, *argumentative completeness*, can preclude learning in hierarchical Bayesian settings, even in the presence of constraints that are designed to promote good epistemic practices.

KEYWORDS

learning, bias, agent modeling, echo chambers, Bayesian learning, philosophical foundations

ACM Reference Format:
Nicholas Asher and Julie Hunter. 2022. When learning becomes impossible. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), June 21–24, 2022, Seoul, Republic of Korea.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3531146.3533078

1 INTRODUCTION

In this paper, we describe and formally analyze a simple epistemic bias, in which people end up being unable to learn, unable to shift their beliefs in the face of evidence. We call this bias *interpretive blindness* (IB). IB is now familiar and all around us—people dying of Covid but refusing to believe that the disease exists despite the evidence, people who believe elections are fraudulent in spite of overwhelming evidence to the contrary, the list goes on. IB is based on a codependence between what beliefs we have and what evidence we use to update those beliefs; it then results from exploiting simple rules of Bayesian inference in a dynamic, iterative process whereby a learner's background beliefs and biases lead her to update her beliefs based on a body of testimony T, while biases inherent in T come back to reinforce her beliefs and her trust in T, further biasing her towards accepting T for future updates. We look at how this codependence affects human agents whose beliefs are guided and shaped by testimony—perhaps the primary way that most people acquire information nowadays. We show that an inability to learn can result, and we show how this inability is enabled and exacerbated by modern machine learning algorithms that can govern what testimony we have access to.

When learning through testimony, an agent acquires beliefs through conversations with other agents, or from books, newspapers or social networks, and so on. Typically, such people lack direct access to the phenomena described via that testimony or cannot analyze the phenomena themselves [Millgram 2015]. Typically too, humans only pay attention to a restricted set of bodies of testimony from a limited number of sources for their information—which makes sense in terms of an agent's limited resources and attention span. These conditions are the fertile ground for the learning inability of IB. In IB, agents' biases preclude learning when an agent tries to exploit new data that are incompatible with or simply distinct from T; agents will discount any evidence that challenges their beliefs. IB is enabled and exacerbated by algorithms that tailor the testimony they provide to that which the agent is already disposed to believe. Our paper formally analyzes the strategic epistemic consequences of these algorithms and of IB. We use Wolpert's 2018 extended Bayesian framework to prove our results.

IB is not only problematic for first order Bayesian approaches but for hierarchical ones [Gelman et al. 2013] as well, because testimony from sources like *Facebook* and other social media, 24/7 media outlets and web interest groups is often *argumentatively complete*, a notion we analyze precisely in Section 4. In an argumentatively complete body of testimony T, the authors of that testimony can respond to and undercut any doubts raised by other data or arguments in a body T' that might threaten T's credibility. A skillful climate denier, for example, will always find a way to undercut the most scientifically careful argument for climate change. Argumentatively complete testimony can undermine higher order constraints and good epistemic practices that guide first order learning.

Our paper starts in Section 2 by introducing the codependence of belief and interpretation relative to testimony and the hypotheses that support it. In Section 3 we formally show how IB can result in first-order Bayesian learning. Section 4 shows how IB can come about in a hierarchical Bayesian learning setting. Section 5 discusses related work while Section 6 develops a game theoretic setting to investigate the complexity of IB. We investigate whether IB is rationally refutable; on certain commonly accepted epistemic assumptions, we conclude that it is not.

2 TESTIMONY AND SOURCES

IB arises in learning because of a codependence between beliefs and the interpretation of evidence, in this case written or linguistically conveyed information. The interpretations we are interested in are judgments about the evidence's trustworthiness. In updating

FAccT '22, June 21–24, 2022, Seoul, Republic of Korea

ACM ISBN 978-1-4503-9352-2/22/06…$15.00
https://doi.org/10.1145/3531146.3533078

our beliefs with new evidence E, our beliefs, particularly about the reliability of E's source, color how we interpret E, how much in short we believe it. But new evidence E' can update our beliefs about the reliability of sources which in turn confer a possibly new degree of belief concerning E. This codependence between evidence and belief dictates how we learn.

Let us look at this codependence more closely in conjunction with a body of testimony. A body of testimony T is a collection of information conveyed by one or more sources that may "promote" or vouch for certain descriptions of events and cast doubt on or disparage others. *The New York Times*, *Fox News*, *CNN*, *Facebook*, *4Chan*, all provide bodies of testimony; their union is also a body of testimony. While such bodies may be consistent or inconsistent, we restrict ourselves here to consistent T. Importantly such bodies are also *dynamic*; they evolve over time as they are updated with new descriptions of events. Dynamic bodies of testimony are ubiquitous in our communicative landscape: on-line, 24/7 news sources as well as particular groups on social media provide continuously evolving, updated coverage of new events. To model this, we shall say that T comes in "stages", where stages can be defined by times or even conversational turns, and each stage T_i is the body of evidence accumulated up to stage i. $T = \{T_1, T_2, ..., T_n, ...\}$ is the collection of all the stages of a dynamic body of evidence. Thus, a body of evidence invites us to iteratively update our beliefs about the trustworthiness of that very body.

Let $\mathcal{T}$ be a collection of (potentially conflicting) bodies of testimony T about some phenomenon P, and assume that a learner $\hat{f}$ does not have independent access to P and can thus only learn about it via $\mathcal{T}$. Learning from $\mathcal{T}$ will require $\hat{f}$ to judge some body of testimony T in $\mathcal{T}$ as credible or trustworthy. Let $\mathcal{H}$ be a set of *evaluation hypotheses*, where each $h \in \mathcal{H}$ evaluates the bodies of testimony T in $\mathcal{T}$. The hypotheses $h \in \mathcal{H}$—background beliefs that may take into account T's source, subject matter, past known accuracy, appeal, and perhaps other elements—define a conditional probability $P(T|h)$ for each $T \in \mathcal{T}$, which we will sometimes write as $h(T)$, where $h(T) = 0$ means T is untrustworthy according to h, and $h(T) = 1$ means T is trustworthy according to h. Following Wolpert's 2018 extended Bayesian framework, $\hat{f}$ determines his belief in T relative to $\mathcal{H}$.

Our learner $\hat{f}$ also has a probability distribution over the evaluation hypotheses in $\mathcal{H}$. Since the bodies of evidence T in $\mathcal{T}$ are dynamic, $\hat{f}$ updates this distribution relative to the stages T_i of T as each T evolves. This is intuitive; testimony T to which $\hat{f}$ attends should be relevant to how $\hat{f}$ updates its beliefs about the trustworthiness of T. The collection $\mathcal{T}$ is also most likely restricted, which is intuitive too: most if not all of us acquire new information from a restricted set of bodies of evidence—a reasonable choice given the balance rational agents need to find between exploiting an already acquired body of evidence and gathering data from other bodies of evidence. And finally, because we restrict ourselves to consistent bodies of evidence T, each T will push a particular point of view on events.

But as the probabilities of the background beliefs, the evaluation hypotheses, are updated over the stages T_i, so too will $\hat{f}$ update her belief in T, using the updated probabilities for her evaluation hypotheses. This codependence can lead to a problem in learning: when we rely on testimony to learn and we restrict the testimony we pay attention to, the confirming evidence for the background beliefs, the evaluation hypotheses and the testimony T they support mutually confirm each other to form a barrier to learning about events that are not mentioned in T. Algorithms of the sort used in social media to build testimony for agents are optimized to continue the themes and ultimately the content of what those agents have previously consulted by exploiting the agent's history of choices that reveal her likes. We will formalize this in terms of a relation between the algorithm's choices and the presence of certain evaluation hypotheses. Such algorithms naturally accelerate the process of what we call below *bias hardening* and IB in the presence of iterated Bayesian updating, despite our learner's rational epistemic practices. We turn to this in the next section.

3 IB IN A FIRST ORDER BAYESIAN SETTING

To formalize IB and its consequences, we first present a simple experiment to show how the codependence of interpretation and belief leads to *bias hardening*.[1] To illustrate, suppose that $\hat{f}$ considers a consistent dynamic body of testimony $T = \{T_1, T_2, ..., T_n, ...\}$ and has two evaluation hypotheses h_1, h_2, where the prior probabilities assigned to h_1 and h_2 by $\hat{f}$ are:

$$P(h_1) = .6, \quad P(h_2) = .4 \tag{1}$$

and the evaluation hypotheses assign probabilities to T as it evolves through stages T_i as follows:

$$P(T_i|h_1) = .8, \quad P(T_i|h_2) = .2 \text{ for all i} \tag{2}$$

We can now calculate the probability of T_1 using the general rule for marginal probabilities in 3. Let $\mathcal{B}$ be $\hat{f}$'s background beliefs; and let the set of all h_i, the alternative hypotheses that are consistent with or assigned non-zero conditional probability relative to $\mathcal{B}$ [L Griffiths et al. 2008; Lampinen and Vehtari 2001; Tenenbaum et al. 2006, 2011], be the set of evaluation hypotheses h_i (so $\{h_1, h_2\}$, in our example).

$$P(x) = \sum_{i=1}^{i=k} P(x|h_i, \mathcal{B}).P(h_i, \mathcal{B}) \tag{3}$$

Then using (1), (2), and (3), we have:

$$P(T_1) = P(T_1|h_1).P(h_1) + P(T_1|h_2).P(h_2) = .56. \tag{4}$$

This is our estimation of our belief in the body of evidence T based on what we have so far. We will continue to update the probability of T given new stages T_i below by distinguishing prior probabilities P_{prior} and updated probabilities P_{post}. Now suppose there is a new conversational turn in T, a new stage of evidence T_2. Given our assumptions, $P(T_2|h_1) = .8$, while $P(T_2|h_2) = .2$, T_2 is supported by h_1 but not by h_2—h_1 and h_2 are consistent with their roles on T_1. Given the dependence of beliefs and interpretation of evidence, T_2 also leads us to re-evaluate our evaluation hypotheses by adapting Bayes' formula to our evaluation hypotheses:

$$P(h_i|T_{n+1}) = \frac{P(T_{n+1}|h_i).P_{prior}(h_i)}{P_{post}(T_n)} \tag{5}$$

[1] [Kelly 2008] describes an informal description of this phenomenon.

Given T_2, whose initial probability we set to what the posterior calculated for T_1—i.e., $P_{post}(T_1) = P_{prior}(T_2)$, we can update our confidence in h_1 as follows:

$$P(h_1|T_2) = \frac{P(T_2|h_1).P_{prior}(h_1)}{P_{post}(T_1)} \approx .86. \qquad (6)$$

Thus, we have posterior probabilities for our evaluation hypotheses as well as for stages of bodies of evidence. The similarly updated probability for h_2 now drops to roughly .14. Using the updated values for h_1 and h_2, we see that T_2, which includes T_1, is now even more believable: $P_{post}(T_2) = .74$. Now suppose that a new bit of evidence, T_3, is added to T. As before, we set $P_{post}(T_2) = P_{prior}(T_3)$. Given our assumptions about our source functions, $P(T_3|h_1) = .8$, we have $P(h_1|T_3) = .96$, while $P(h_2|T_3) \approx 0.04$, and confidence in T_3 is also updated: $P_{post}(T_3) = .776 \approx .78$. Updating h_1's probability conditional on new evidence T_4 now yields a value of $.989 \approx .99$, while $P(h_2|T_4) = 0.008 \approx 0.01$. By the time we get to T_5, the probability of h_1 will have gone to 1, while $P(h_2) = 0$, and $P(T_5) = .8$. In sum, as n increases, the updated probabilities of h_1 go to 1 and $P(T_n) \rightarrow P(T|h_1)$, that is, to the strength of h_1's support for T.

Our codependence of belief and evidence suggests a loopy structure (cyclic graph) for updating. However, by exploiting stages, we can disentangle such structures; and efficient approximations are possible in disentangled structures [Murphy et al. 2013]. Proposition 1 below shows a convergence under certain assumptions. Let $\mathrm{P}_n(h_i)$ be the probability of h_i after conditionalizing on T_n and $\mathrm{P}_n(T)$ the value of T after n conditional updates as defined above.

Let's now move to a more general setting. Let $\hat{f}$'s evaluation hypotheses $\mathcal{H}_{\hat{f}}$, come with a probability distribution. An agent could have, among the many evaluation hypotheses that she countenances, an evaluation hypothesis h for which the conditional probability of T given h increases as T evolves. The support for T might increase (or decrease) as T gets more extended with more and more stages.

Definition 1. *An evaluation hypothesis $h \in \mathcal{H}_{\hat{f}}$ is positive sensitive to $T = \{T_1, T_2, ...\}$ iff $P(T_n|h) > .5$ and is monotone increasing for all n.*

How bodies of testimony are constituted for ordinary learners is not always clear. Learners can assemble their own body of testimony, and typically, they must concentrate on some testimony to the exclusion of other testimony. What are the criteria? Well, sometimes other actors can guide the acquisition of testimony. What we shall call *Facebook-like algorithms* from our epistemic perspective are a way social media and news organizations can steer learners to a certain body of testimony. Their role is to bring testimony to an agent's attention that feeds and updates, in fact constructs, an evaluation hypothesis that keeps the learner coming back to the same type of information, often the same set of sources of information. More formally, the role of such an algorithm is to construct a positive sensitive evaluation hypothesis in $\mathcal{H}_{\hat{f}}$.

Definition 2. *A* Facebook-like *(FB) algorithm g for $\hat{f}$ constructs a body of testimony from a set of bodies of testimony $\mathcal{T}$ that $\hat{f}$ uses to update her hypotheses and $g(\mathcal{T}) = T = \{T_1, T_2, ...\}$ with $g_n(\mathcal{T}) = T_n$ iff there is an $h \in \mathcal{H}_{\hat{f}}$ such that: (i) $\exists n\, P(g_n(\mathcal{T})|h) > .5$ and (ii) $\forall m > n\, \exists k\, (P(g_{k+m}(\mathcal{T})|h) > P(g_m(\mathcal{T})|h)$ as long as $P(g_{k+m}(\mathcal{T})|h) \neq 1)$.*

An FB algorithm $g_n(\mathcal{T})$ at each stage n provides information that does not decrease h's support for T. Thus g makes h positive sensitive to T through a choice of stages T_n. In addition, however, g will eventually keep on increasing the support of h for $g(\mathcal{T})$. We assume that if g is an FB algorithm for $\hat{f}$, then $\hat{f}$ updates her hypotheses based on the testimony fed by g. We suspect that actual social media algorithms are FB though we do not have a proof of this. The possibility, however, seems to us very real.

Proposition 1. *Suppose testimony $T = \{T_1, T_2, ..., T_n, ...\}$, and with $h_1 \in \mathcal{H}$ positive sensitive to T and with $P(h_1) \neq 0$, while $P(T_n|h_j) < .5$ and is monotone decreasing for all n and for all $h_j \in \mathcal{H}, h_j \neq h_1$. Then:*

$$As\ n \rightarrow \infty,\ \mathrm{P}_n(T) \rightarrow limsup(P(T_n|h_1)),$$
$$\mathrm{P}_n(h_1) \rightarrow 1\ and\ \mathrm{P}_n(h_j) \rightarrow 0\ for\ j \neq 1$$

Given the calculations above and using standard updating rules for the probabilities P assigned by $\hat{f}$, if $P(T_i|h_1)$ is monotonic increasing with respect to i and $P(T_i|h_j)$ for any $j \neq 1$ is monotonic decreasing, then the updates of $P(T_i)$, $P(h_1|T_i)$ and $P(h_j|T_i)$ will follow the pattern of our experiment above and converge to the support of h_1, 1, and 0 respectively. □

Corollary 1. *If $\hat{f}$ uses an FB algorithm and $h_j, j \neq 1$ are as in Proposition 1*

$$As\ n \rightarrow \infty,\ \mathrm{P}_n(T) \rightarrow limsup(P(T_n|h_1)),$$
$$\mathrm{P}_n(h_1) \rightarrow 1\ and\ \mathrm{P}_n(h_j) \rightarrow 0\ for\ j \neq 1.$$

Given that an FB algorithm entails the existence of a positive sensitive hypothesis h_1 to T and that h_1 must have non 0 probability, the result follows from Proposition 1 □.

We now introduce three important properties of evaluation hypotheses.

Definition 3. *An evaluation hypothesis h for a set of bodies of testimony $\mathcal{T}$ is* consistent *iff for $T, T' \in \mathcal{T}$, if $T \cup T'$ is inconsistent, then $P(T|h) = 1 - P(T'|h)$. An evaluation hypothesis h is* probability-wise model complete (PWMC) *for T and some topic t iff: for any putative piece of evidence ϕ on t if for no stage T_i $T_i \models \phi$ (ϕ is not predicted or included in any stage of T), then $P(\phi|h) = 1 - P(T|h)$.*

Definition 4. *An evaluation hypothesis $h \in \mathcal{H}$ makes T* potentially trustworthy *($h \models T$), if as $n \rightarrow \infty$, $P(T_n|h) \rightarrow 1$.*

Proposition 2. *Suppose $g(\mathcal{T}) = T$ and g is an FB algorithm for $\hat{f}$. Then there is an evaluation hypothesis h such that $h \models T$.*

Let g be an *FB* algorithm for $\hat{f}$. Then as we showed above, there is an h that is positive sensitive to T and by the property (ii) of Facebook algorithms, $lim_{n\rightarrow\infty}P(T_n|h) = 1$. □
While consistency seems a basic requirement of evaluation hypotheses, a potentially trustworthy evaluation hypothesis is a kind of "soundness" or accuracy assumption about a body of evidence. For an agent who remains wedded to a body of testimony, such an assumption also seems rational.

PWMC hypotheses generalize consistent hypotheses, but what is their rationale? FB algorithms by themselves don't lead to PWMC. But they do facilitate it. As g feeds testimony to $\hat{f}$ that lends more and more support to an eventually trustworthy hypothesis h, a natural thought is for $\hat{f}$ to assume that T_{i+1} provides a more complete coverage of the pertinent facts than T_i. So elements ϕ that are not mentioned in any T_i or are in some way incompatible with T are either not relevant, or just false. The PWMC condition codifies this for evaluation hypotheses in terms of an operation akin to negation as failure in Prolog; if h makes T probability wise model complete, then if given some newsworthy and relevant topic t and$T$ doesn't mention ϕ, then h supports $\neg\phi$ to the extent that h supports T. For simplicity, we will assume that the topic parameter is constant in our discussions below and drop it from the formalism.

Proposition 3. *Let $\mathcal{T}$ be a set of consistent bodies of testimony and let g be a FB algorithm for $\hat{f}$ with $g(\mathcal{T}) = T$ with $h_1 \in \mathcal{H}_{\hat{f}}$ the entailed positive sensitive hypothesis to T. Let the priors on $h_i \in \mathcal{H}, h_i \neq h_1$ be as in Proposition 1 Then:*

$$As\, n \to \infty,\ \mathsf{P}_n(T) \to 1. \tag{1}$$

Suppose in addition, h_1 is PWMC for T and $T \not\models T'$. Then:

$$As\, n \to \infty, \mathsf{P}_n(T') \to 0 \tag{2}$$

To show (1), note that $P(T_1|h_1) > .5$ and since as $n \to \infty$, $\mathsf{P}_n(h_1) \to 1$, after a certain point $P(T_n|h_1)$ is monotone increasing. Then by Proposition 1, $\mathsf{P}_n(T) \to P(T_n|h_1)$. Since h_1 makes T potentially trustworthy, as $n \to \infty, \mathsf{P}_n(T) = 1$. To show (2), suppose h_1 is PWMC for T. Given that $T \not\models T'$, $h_i(T_i') = 1 - h_i(T_i)$ for each i, and the expected probability of T' will decrease strictly monotonically over n, as $\mathsf{P}_n(h_1) \to 1$. So as $n \to \infty,\ \mathsf{P}_n(T') = 0$. □

Corollary 2. *Let the priors on $h_i \in \mathcal{H}, h_i \neq h_1$ be as in Proposition 1. Suppose in addition, $T \models \neg T'$, then:*

$$As\, n \to \infty, \mathsf{P}_n(T') \to 0 \tag{2}$$

Note that our agent may have many evaluation hypotheses and the result of Proposition 3 is unchanged. Crucially $\hat{f}$ has updated his beliefs only on T due to g. This matches our intuitions about what agents actually do. As long as the codependence between background beliefs and bodies of evidence holds and certain bodies of evidence are supported more than others, belief in some bodies of evidence $T \in \mathcal{T}$ will be strengthened, while belief in bodies of evidence in conflict with T or just different from T will be weakened. Importantly, this can happen *merely by T_i repeating content already in T_k* for $i > k$. Such repetitions of content are commonplace on social media sites and news sites that broadcast continuously. In addition, the assumption of a PWMC evaluation hypothesis is rather mild; it reflects an agent's mistrust of bodies of evidence other than the ones he relies on—a rather common situation.

Proposition 3 impacts the marginalization of new data, because if its assumptions are met, as $\mathsf{P}_n(T') \to 0$, $\hat{f}$ discounts evidence from T', despite the presence of evaluation hypotheses supporting T'.

Proposition 4. *Suppose evidence ϕ such that $T' \models \phi$, $T \not\models \phi$ and T, T', and $\hat{f}$'s evaluation hypotheses are as in Proposition 3 and $\hat{f}$ conforms to Bayesian learning. Then:*

$$As\, n \to \infty,\ \mathsf{P}_n(\phi) \to 0.$$

Since $\hat{f}$ conforms to Bayesian learning, the marginal probability for ϕ is based on Equation 3 and the set of hypotheses h_i in Equation 3 is the set $\mathcal{H}$ that for $\hat{f}$ pronounce on testimony that mentions or asserts ϕ. By Proposition 1, as $n \to \infty$, $\mathsf{P}_n(h_1) \to 1$. By Proposition 3, $\mathsf{P}_n((T'h_1) \to 0$. But for all other h_k such that $h_k(T') \neq 0$, by Proposition 1 again, as $n \to \infty$, $\mathsf{P}_n(h_k) \to 0$. But then $\mathsf{P}_n(\phi|h_i, \mathcal{B}) \to 0$ for all relevant h_i. Given Equation 3, the result follows. □

In this situation, $\hat{f}$ assigns no credence to ϕ. The prior beliefs of $\hat{f}$ may so limit the alternative hypotheses h_i such that even an actual fact ϕ will have a marginal probability of 0; $\hat{f}$ will discount ϕ completely.

Now consider general learning in this situation, defined in Wolpert's 2018 extended Bayesian framework via Bayes's formula.

$$P(h|x, \mathcal{B}) = \frac{P(x|h, \mathcal{B}).P(h|\mathcal{B})}{\sum_{i=1}^{i=k} P(x|h_i, \mathcal{B}).P(h_i|\mathcal{B})} \tag{7}$$

To learn a hypothesis ψ about some event e, $\hat{f}$'s estimation of ψ at some stage $P_n(\psi)$ based on her evidence should be closer to the objective or ideal assignment (posterior) ψ_p, than to her prior probability for ψ, $P_0(\psi)$. Similarly for marginal probabilities: $\mathsf{P}_n(x)$ should track x_p, the posterior of x, given a random sampling of X. We consider loss functions $\mathcal{L}(\mathsf{P}_n(\psi), \psi_p)$ and $\mathcal{L}(\mathsf{P}_n(x), x_p)$. The greater divergence between the ideal posterior probability and the Bayesian subjective estimation of that probability, the worse will be the score for $\hat{f}$'s learning. We say that $\hat{f}$ cannot learn ψ if her updates do not eventually decrease loss; i.e. we cannot show $lim_{n\to\infty}\mathcal{L}(\mathsf{P}_n(\psi), \psi_p) < \mathcal{L}(\mathsf{P}_0(\psi), \psi_p)$.

Proposition 5. *Suppose $\hat{f}$ is a Bayesian learner with evaluation hypotheses and testimony T, T' as in Proposition 3. Let ψ be a new hypothesis with all evidence e confirming ψ such that $T' \models e$. Then $\hat{f}$ is incapable of learning ψ.*

Consider e such that $T' \models e$ and e confirms ψ. So the true posterior $P_p(\psi|e) > P(\psi)$, with $P(\psi)$ the prior on ψ. Suppose $\hat{f}$'s evaluation hypotheses and probabilities have been updated via T as in Proposition 3 and let that give the "prior" probability for the new hypothesis ψ. By Proposition 4, as $n \to \infty,\ \mathsf{P}_n(e) \to 0$. In the limit, Bayesian learning as specified by equation (4) simply isn't defined when $\mathsf{P}_n(e) = 0$. So assuming e is discounted as evidence in updating, we set $P(\psi|e, T_n) = P(\psi|T_n)$. But this is just $P_0(\psi)$, $\hat{f}$'s prior on ψ. It follows that as $lim_{n\to\infty}\ \mathcal{L}(\mathsf{P}_n(\psi), \psi_p) \not< \mathcal{L}(\mathsf{P}_0(\psi), \psi_p)$. □

Proposition 5 is a formal statement of IB in a first order setting. It shows that under certain conditions, $\hat{f}$ will be incapable of learning any hypothesis that involves a dependence on testimony not in T, upon which $\hat{f}$ has formed his beliefs. $\hat{f}$ is interpretively blind to any possibilities outside of T.

4 IB IN HIERARCHICAL BAYESIAN LEARNING

It's not unreasonable to rule out new evidence from unreliable testimony, *provided* the assignment of one's evaluation hypotheses to the testimony is reasonable. But nothing in our discussion above forces the evaluation hypotheses to be be reasonable. Without any constraints, $\hat{f}$'s evaluation hypotheses may rule out evidence that is completely grounded in reality and comes from testimony that an ideal rational agent would trust.

To solve this problem, we need to correct the background beliefs $\mathcal{B}$. Ideally, a rational agent should control for the biases in testimony by consulting several different bodies of testimony. However, $\mathcal{B}$ cannot be corrected itself by evidence, because that evidence is already discounted if it conflicts with $\mathcal{B}$. Very clearly, background beliefs can be a source of bad epistemic biases, and they can prevent straightforward corrections to improve one's beliefs as Bayesian learning would have us do.

Hierarchical Bayesian models were designed to address this problem [Gelman et al. 2013]. In hierarchical Bayesian models, a Bayesian learning model like the one we have discussed in Section 3 has certain parameters; the one parameter we have is our evaluation hypotheses providing the reliability of testimony. At a second level of the hierarchy, we could have a Bayesian learning model concerning evaluation hypotheses, in which we could detail factors that would allow us to estimate reliably the accuracy of an evaluation hypothesis. Abstractly, we would have evaluation hypotheses about evaluation hypotheses that would discuss factors like the consistency or the predictive accuracy of a testimony source, or the extent to which testimony from other sources agrees with its content. One could also require a longer or more thorough exploration of the data about the phenomenon before the agent's restricting himself to a small subset for exploitation (once again an application of the work in [Cesa-Bianchi and Lugosi 2006]). All of these ideas and more have been proposed.

Simply requiring evaluation hypotheses that obey exogenous constraints, however, begs the question of why $\hat{f}$ should accept them. In fact, the interdependence of testimony, new information and background beliefs can make the resort to higher order parameters to resolve IB a failure because a body of dynamic testimony T, when directed by a conversational agent for the purposes of persuading and keeping his audience, can react to and attack not only a conflicting body of testimony T' but also sources supporting it. This behavior provides arguments for or against not only first order evaluation hypotheses, as we've seen with the notion of consistency, but also for higher order functions and in fact sequences of evaluation hypotheses.

To formalize this picture, we assume a hierarchy of sets of evaluation hypotheses where,

$$h^{n+1} : h^n \rightarrow [0, 1], \text{ for } h^{n+1} \in \mathcal{H}^{n+1}.$$

Hypotheses at level $n + 1$ are related to evaluation hypotheses at level n via *rationality.*

Definition 5. *A set of sets of evaluation hypotheses* $\mathcal{H} = \{\mathcal{H}^1, \mathcal{H}^2, ..., \mathcal{H}^n\}$ *is* rational *iff for all* $m < n$, $h^m_k \in \mathcal{H}^m$, $P(h^m_k) = \lambda \sum_{h_j \in \mathcal{H}^{m+1}} P(h^{m+1}_j).P(h^m_k | h^{m+1}_j)$ *for some normalizing factor* λ.

A rational set of sets of evaluation hypotheses is one in which the probability of evaluation hypotheses at one level reflects what higher levels say about it. We will assume rationality of $\mathcal{H}$.

Given rational $\mathcal{H} = \{\mathcal{H}^1, \mathcal{H}^2, ..., \mathcal{H}^n\}$, we now lift our notions of support to sequences. We define a $\mathcal{H}^n$ sequence $\sigma \in \prod^n_{i=1} \mathcal{H}^i$ of consistent evaluation hypotheses to support T ($\sigma \mid\!\approx T$) (or that make T potentially trustworthy—$\sigma \models T$) iff the $\mathcal{H}^1$ element h^1_σ of σ is positive sensitive to T (makes T potentially trustworthy) and every element of σ has non-0 probability given $\mathcal{H}$. Conversely, we say that $T \mid\!\approx \sigma$ iff for each element h^i_σ of σ $P(h^i_\sigma | T_j)$ is eventually monotone increasing for all stages T_j. We note that $\sigma \mid\!\approx T \rightarrow T \mid\!\approx \sigma$.

Let σ^k be the subsequence of σ such that $\sigma^k = \sigma \upharpoonright (\prod^n_{i=k} \mathcal{H}^i)$. For $h^1 \in \mathcal{H}^1, \sigma^2(h^1)$ signifies the support h_1 receives from the higher order functions in σ via Definition 5.

Definition 6. *An* $\mathcal{H}^n$ *sequence* σ_1 undercuts T *iff for any* $\sigma^i \in \mathcal{H}^i$, *if* $\sigma^i \mid\!\approx T$, *then* $\sigma^{i+1}_1(h^i_\sigma) = 1 - P(T | h^1_\sigma)$

Definition 7. ϕ disagrees *with* T' *just in case* $P(T|\phi) < P(T)$.

Definition 8. *Given* $\mathcal{H} = \{\mathcal{H}^1, \mathcal{H}^2, ..., \mathcal{H}^n\}$ T attacks T' *iff (i) if* $\sigma \mid\!\approx T$, *then* $P(T' | h^1_\sigma) = 1 - P(T | h^1_\sigma)$ *and (ii) for any* $\mathcal{H}^m$ *sequence* σ^m, $m < n$ *if* $\sigma^m \mid\!\approx T'$, $\exists h^{m+1} \in \mathcal{H}^{m+1}$ *such that* $(P(h^{m+1}|T) > .5$ *and* $h^{m+1}(\sigma) = 0)$.

Definition 9. T *is* argumentatively complete *iff:*
(i) $(T' \models \phi$ *and* $\text{Disagree}(\phi, T)) \rightarrow \text{Attack}(T, T')$*; (ii) If* $T_n \not\models \phi$ *but* $P(T_n|\phi) \geq P(T_n)$, *then* $T_{n+1} \models \phi$. *(iii) for any T undercutting sequence* σ^m, $\exists h^{m+1} \in \mathcal{H}^{m+1}$ *such that* $(P(h^{m+1}|T) > .5, h^{m+1}(h^m_\sigma) = 0)$ *(iv)* $\exists \mathcal{H}^n$ *sequence* σ *such that* $\sigma \models T$.

Proposition 6. *If T is argumentatively complete, then there is an h such that h is PWMC for T.*

Assume that T is argumentatively complete. Then $\exists \sigma \in \mathcal{H}^n$ *such that* $\sigma \models T$. Since $\sigma \models T$, $\sigma \mid\!\approx T$. Now assume $T_n \not\models \phi$ for some ϕ for all stages n. But then $P(T_n|\phi) < P(T_n)$ for each stage T_n of T. But then T and ϕ disagree and so T attacks ϕ. By the definition of attack, $P(\phi|h^1_\sigma) = 1 - P(T|h^1_\sigma)$. So h^1_σ is PWMC. □

Proposition 7. *Let T be argumentatively complete with a rational set of evaluation hypotheses* $\mathcal{H}$ *with* $\sum_{h^1 \in \mathcal{H}^1} P(h^1) \neq 0$ *and probabilities updated on T.*

$$\text{As } n \rightarrow \infty,\ \mathsf{P}_n(T) \rightarrow 1. \tag{1}$$

In addition suppose there is a $T' \nsubseteq T$.

$$\text{As } n \rightarrow \infty,\ \mathsf{P}_n(T') \rightarrow 0. \tag{2}$$

We first show (1). Since T is argumentatively complete, $\exists \sigma \in \mathcal{H}$ such that $h^1_\sigma \models T$. We need to show that for some such h^1_σ, $P(h^1_\sigma) \neq 0$ relative to $\mathcal{H}$. Suppose that $P(h^1_\sigma) = 0$, for all h^1_σ such that $h^1_\sigma \models T$. By rationality, for each such h^1_σ, $P(h^1_\sigma) = \lambda \sum_{h_j \in \mathcal{H}^2} P(h^2_j).P(h^1_\sigma | h^2_j) = 0$. Thus, all the non-0 probability mass of $\mathcal{H}$ falls on T undercutting sequences σ_i. But for each such T undercutting σ_i of length m, since T is argumentatively complete, there is an evaluation hypothesis h^{m+1} supported by T such that $P(\sigma_i | h^{m+1}) = 0$. Since $\mathcal{H}$ has only finitely many levels, at some level k all T undercutting sequences σ_j get 0 probability. This, together with the fact that $\sum_{h^1 \in \mathcal{H}^1} P(h^1) \neq 0$, contradicts the assumption that $P(h^1_\sigma) = 0$. Since T is argumentatively complete, any sequence supporting any h^1 where $P(T|h^1) < P(T)$ will eventually get probability 0; so $\sum_{\{h^1 : P(T|h^1) \geq P(T)\}} P(h^1) = \sum_{h^1 \in \mathcal{H}^1} P(h^1)$. Moreover, as P_n gets updated, as $n \rightarrow \infty$, the h^1 such that $\mathsf{P}_n(T|h^1) \geq \mathsf{P}_n(T)$ turn out to be such that $h^1 \models T$. The conditions on first order evaluation hypotheses in $\mathcal{H}$ of Proposition 1 are now met. By Propositions 1 and 3, as $n \rightarrow \infty$, $\mathsf{P}_n(h^1_\sigma) \rightarrow 1, \mathsf{P}_n(h^1_i) \rightarrow 0$ for $i \neq 1$. By Proposition 3, $\mathsf{P}_n(T) \rightarrow 1$.

To show (2), by Proposition 6, h^1_σ is also PWMC for T. As $n \rightarrow \infty$, *since* $\mathsf{P}_n(h^1_\sigma) \rightarrow 1, \mathsf{P}_n(T') \rightarrow 0$. □

Proposition 8. *Suppose T is argumentatively complete. Let $\hat{f}$ be a hierarchical Bayesian learner whose evaluation hypotheses are rational and are updated on T. If $T' \subsetneq T$ such confirms a hypothesis h that T does not, then $\hat{f}$ is incapable of learning h.*

Claim 2 of Proposition 7 shows that $\mathsf{P}_n(T') \rightarrow 0$. Then apply Proposition 5. □

Argumentatively complete testimony reduces the case of higher order Bayesian frameworks to our first order setting for IB. What is troubling about IB is that our learner $\hat{f}$ may hold onto an argumentatively complete T regardless of how inadequate it is in the eyes of others or standard epistemic criteria; an argumentatively complete theory will always eventually find a reply to any attack or any doubt $\hat{f}$ might acquire.

Argumentatively complete testimony isn't just an abstract concept; many social media and news sites already approximate this condition. Outlets like *NewsMax* or *One Amercan News Network* that have a particular political bias will attack the credibility of stories from other bodies of testimony that have gone against a narrative they were and are promoting; darker conspiracy spinning websites like those promoting QAnon will attack arguments against their theories once they become aware of them.[2] In anecdotal support of our claims, consider Michelle Goldberg's "It's Marjorie Taylor Greene's Party Now" *New York Times*, 2/2/2021) description of a group in IB: "American conservatism — particularly its evangelical strain — has fostered derangement in its ranks for decades, insisting that no source of information outside its own self-reinforcing ideological bubble is trustworthy."

A crucial component of argumentatively complete testimony T is that it promotes evaluation hypotheses that both make T eventually trustworthy but also PWMC for T. Sources like *the New York Times* embody this in their slogan *all the news that's fit to print*, but there's a commercial reason for this outcome; news sites and social media are out to capture market share and so they naturally promote themselves as accurate and complete at least in a certain domain. The nature of contemporary testimony leads agents naturally to a situation where IB occurs.

But how does argumentative completeness relate to FB algorithms? We suspect that FB algorithms don't entail argumentative completeness, but they facilitate it like they do for PWMC. An FB algorithm g's choosing testimony T_j for $\hat{f}$ should ideally detract from other possible additions to $\hat{f}$'s evidence that might lead $\hat{f}$ to find other testimony besides that provided by g. Since the goal of providing g is to reinforce T and the evaluation hypotheses supporting it, a possible consequence is for g to provide testimony that not only supports T but attacks any evidence that might detract from T.

5 COMPARISONS TO PRIOR WORK

IB is an epistemological bias that we briefly described in a preliminary abstract [Asher and Hunter 2021]. IB is clearly related to confirmation bias [Lord et al. 1979; Nickerson 1998; Oswald and Grosjean 2004], in which agents interpret new evidence in a way that confirms their beliefs, and to the framing biases of [Tversky and Kahneman 1975, 1985]. People tend to see in the evidence what they believe. These forms of bias, however, concern how beliefs and bias influence interpretation, painting only part of the picture of IB. [Asher and Paul 2018] shows a codependence between beliefs and the interpretation of ambiguous or underspecified elements in a text and postulates a similar circular structure to that which we have exploited for belief and interpretation of evidence in analyzing IB. Further, unlike much of the psychological and philosophical literature which either claims that biases like IB arise from bad epistemic practices or aren't really beliefs at all, or finds epistemologically exogenous justifications for it [Cassam 2016; Dardenne and Leyens 1995; Ichino and Räikkä 2020], we show how IB is a natural outcome of Bayesian updating, rational resource management and the belief interpretation codependence.

IB is also related to what [Jamieson and Cappella 2008; Nguyen 2020] have called echo chambers and epistemic bubbles. Epistemic bubbles are the sort of epistemic structures that result from Proposition 1. Contrary to what [Nguyen 2020] claims, we show that they are not so easy to get rid of. Once the potentially trustworthy hypothesis is sufficiently entrenched, which can happen very quickly as we show in Section 3, simply bringing evidence inconsistent with an accepted body of evidence will not liberate the learner from his IB predicament. Evidence inconsistent with the primary source of our learner's beliefs will be simply rejected, as we show formally in Corollary 2. Following [Jamieson and Cappella 2008], [Nguyen 2020] claims that echo chambers are much more problematic epistemically. We have shown this rigorously via Proposition 3 as well as in Propositions in Section 4. As we show below in Section 6 (see Proposition 10), there is no escape from these echo chambers, once certain epistemic principles are admitted to. We have also shown conceptual and formal links between epistemic bubbles and echo chambers. In fact epistemic bubbles lead naturally to echo chambers, as we show in Sections 3 and 4.

IB is also related to so called *epistemic bootstrapping* [Douven and Kelp 2013; Vogel 2008; Weisberg 2010]. Epistemic bootstrapping is a phenomenon in which an agent A exploits a hypothesis h she is interested in confirming in the very process of confirming h. IB is somewhat different. In IB, A has a hypothesis h that confers a certain probability on evidence E. Suppose h assigns a relatively high probability to E. As more evidence E' comes in, however, and E' tends to confirm E, then A's confidence in h should increase. And A doesn't just exploit one hypothesis but other hypotheses that may be opposed and that would tend to disconconfirm h. As [Douven and Kelp 2013] argues, bootstrapping is not in itself bad; in fact this is just what scientists do. What matters is whether the process involved in confirming h could disconfirm h. So in principle, the inductive practices that lead to IB aren't easily criticizable as instances of bad epistemic bootstrapping. Most likely unbeknownst to A, however, the epistemic deck is stacked against her, when an FB algorithm is used to feed her testimony. In principle this is quite a different situation epistemically from a case of bad epistemic bootstrapping.

One can also see IB as a concrete application of work on determining an optimal allocation of resources to the exploration and exploitation of sources [Auer et al. 2002; Banks and Sundaram 1994; Burnetas and Katehakis 1997; Cesa-Bianchi and Lugosi 2006; Garivier and Cappé 2011; Lai and Robbins 1985; Whittle 1980]. It is also

[2] See Stuart A. Thompson, "Three Weeks Inside a Pro-Trump QAnon Chat Room" *NY Times*, Jan 26, 2021).

related to work on generalization in machine learning. Epistemic biases affect generalization and learning capacity in ways that are still not fully understood [Kawaguchi et al. 2017; Lampinen and Vehtari 2001; Neyshabur et al. 2017; Zhang et al. 2016]. [Zhang et al. 2016] show that standard techniques in machine learning for promoting good epistemic biases and generalization—training error minimization, regularization techniques like weight decay or dropout, or complexity measures used to minimize generalization error (the difference between training error and test error)—do not necessarily lead to good generalization and test performance. Argumentatively complete testimony T incorporates an adversarial attack mechanism against any good epistemic practices that might discount T. It’s this mechanism that guarantees IB.

The argumentation literature [Amgoud and Demolombe 2014; Dung 1995] is also relevant to IB. If testimony T is *argumentatively complete*, then T always provides a counterargument to an attack against T–much like an acceptable argument in [Dung 1995]. In addition, however, an argumentatively complete T also supports higher order evaluation hypotheses that support hypotheses that support T. There are also important connections to the literature on trust [Castelfranchi and Falcone 2010]; in our set up learning agents trust certain sources over others, and our higher order setting invokes a hierarchy of reasons. Nevertheless, the argumentation and trust-based work of which we are aware is complementary to our approach. An argumentation framework takes a possibly inconsistent belief base and imposes a static constraint on inference in such a setting. Similarly, trust is typically modeled in some sort of static modal framework. By contrast, ME learning games and the whole Bayesian framework are dynamic, with beliefs evolving under evidence and game strategies evolving under agent interaction. It is this dynamic evolution that is crucial to our approach and, we think, to modeling agents and learning. In sum, we are not looking at the problem of consistency, but rather the problems of entrenchment and bias.

6 THE COMPLEXITY OF IB

IB is a result about learning. IB is a suboptimal but natural outcome of the way contemporary bodies of evidence are set up and how humans interpret them. Given our set up, everything turns on what body of evidence on which to update and with which evaluation hypotheses.

If IB is suboptimal, its effects are still more worrisome, because agents in the grip of IB are often unwilling or incapable of changing their beliefs so as to be able to learn. Of course, our learner might just be happy with T; perhaps he needs no more accurate or more truthful body of testimony. He may not be interested in learning anything beyond what T presents him with. In this section, however, we assume a learner who might be interested in learning but has difficulting escaping his IB prison. We assume a rational learner $\hat{f}$ who updates according to his evaluation hypotheses; so if he has an evaluation hypothesis that confers a high probability on some T, he will update on T. We’ve seen that $\hat{f}$ can get IB when he unduly restricts the bodies of evidence which serve as the basis of update or when he attends to an argumentatively complete testimony. So key to removing IB is to get $\hat{f}$ to change his hypotheses and consider other evidence that that to which he is wedded.

Anecdotally, we have a lot of evidence that IB is hard to escape[3] In general, however, we lack a precise analysis of its difficulty. In this section, we introduce a game theoretic method that shows IB is not only hard to defeat but it can even be hard to detect (leading to self-deception). We will see that the choice of epistemic paradigms is important.

To motivate our approach, consider how an actual conversation might go between our learner $\hat{f}$ in the grip of IB and a person E who wants to correct his problem. E might question $\hat{f}$’s reasons for believing some proposition ϕ; she might try getting $\hat{f}$ to consider different bodies of evidence T' that might disconfirm ϕ. $\hat{f}$ might accept T' or he might argue against it—by providing, for example, reasons why T' is not trustworthy or why the arguments supporting T' are faulty. E might attack those arguments or provide new evaluation hypotheses for consideration. Our ME games formalize this interaction.

In an ME learning game $\mathcal{G} = (V^{\infty}, Win)$, the two players, our investigator E and our Bayesian learner $\hat{f}$, construct a larger "conversation" by consecutively playing finite strings from the vocabulary V. Win specifies the winning condition of E. The vocabulary V of an ME learning game $\mathcal{G}$ consists of sequences of evaluation hypotheses (with some abuse of notation, we’ll take a single h^n_j to be a one place sequence) and a predicate ACCEPT. ACCEPT means that $\hat{f}$ accepts the last suggestion by E and confers upon it a non zero probability mass. Our ME learning games are subject to several constraints.

A. *Knowledge first* [Williamson 2002]: this is a constraint from formal epistemology; $\hat{f}$ only adds a sequence σ to $\prod_1^n \mathcal{H}^i_{\hat{f}}$ for $\mathcal{H}^i_{\hat{f}}$ i-th level evaluation hypotheses in $\mathcal{H}_{\hat{f}}$ if he has no argument that attacks σ—in other words no evaluation hypothesis $h^{n+1} \in \mathcal{H}^{n+1}_{\hat{f}}$ such that $h^{n+1}(\sigma) = 0$.
B. The Jury in an ME learning game is epistemologically competent; i.e. it sanctions only evaluation hypotheses that advance learning.
C. E may only add sequences of evaluation hypotheses sanctioned by the Jury. We assume this to be a finite set $\mathcal{H}_J$.
D. Both players may only propose consistent and rational sequences.
E. $\hat{f}$ has learned from some body of evidence T, which is common knowledge.
F. $\hat{f}$ may only not accept a proposal σ of E, if he has a reason to do so—i.e., if he has an evaluation hypothesis $h^{n+1} \in \mathcal{H}^{n+1}_{\hat{f}}$ such that $h^{n+1}(\sigma) = 0$.

We say a sequence $\sigma \in \prod_1^n \mathcal{H}^i$ to be positive if for each element h^{m+1} and h^m of σ $h^{m+1}(h^m) >> 0$. A sequence σ nullifies a sequence σ_1, if for all m and for h^m_1 of σ_1, the h^{m+1} of σ is such that $h^{m+1}(h^m_1) = 0$. We can have two sequences each one nullifying the other. This formally represents an n round argument, with each round $j + 1$ offering a counterargument to the argument of round j. We will say that a hypothesis h^1 is T *positive* if h^1 is positive and $P(T|h^1) = 1$

[3] See Thompson, cited in note 2.

We now define the moves of a game $\mathcal{G}$, in which we suppose a body of evidence T that $\hat{f}$ has attended to and a body of evidence T' inconsistent with T. E plays first, then $\hat{f}$ then replies. The game ends if $\hat{f}$ plays ACCEPT, which implies that he adds a hypothesis h^1_* to $\mathcal{H}^1_{\hat{f}}$, with a non-0 probability mass and with with high $P(T'|h^1_*)$, where $T' \cup T$ is inconsistent.

(m1) E proposes T'-positive $h^1 \in \mathcal{H}^1_J$ to be added to $\mathcal{H}^1_{\hat{f}}$.

(m2) Suppose at round $k \geq 1$ of ρ in $\mathcal{G}$ E has proposed a T' positive h^1. At $k + 1$ $\hat{f}$ may play ACCEPT.

(m3) Suppose at round k of ρ in $\mathcal{G}$ E has proposed a a T' positive h^1. At $k + 1$ $\hat{f}$ may play a nullifying $h^2 \in \mathcal{H}^1_{\hat{f}}$ such that $h^2(h^1) = 0$, if there exists such $h^2 \in \mathcal{H}^2_{\hat{f}}$.

(m4) Suppose E has proposed a positive sequence σ of length m and with h^1_σ T' positive at round k of ρ in $\mathcal{G}$. At round $k + 1$ $\hat{f}$ may respond with sequence of length $m + 1$ nullifying σ.

(m5) Suppose at round k of ρ of $\mathcal{G}$, $\hat{f}$ has proposed an m-length sequence σ nullifying a positive σ_* proposed by E, with T' positive $h^1_{\sigma_*}$. E may respond at round $k+1$ of ρ with a positive $m + 1$ length sequence $h^{m+1}_*.\sigma_*$, with $h^{m+1}_1(h^m) \neq 1$ for h^m in σ.

(m6) Suppose at round k of ρ in $\mathcal{G}$, E has proposed a positive sequence σ of length m and with h^1_σ T' positive. At round $k + 1$ $\hat{f}$ may play ACCEPT, which implies that he adds σ to $\prod^n_1 \mathcal{H}^i_{\hat{f}}$.

We note that if move (m6) occurs, $\hat{f}$ assigns h^1_* and T' a non-0 probability mass and updates with evidence T', which makes the ACCEPT move coherent.

Suppose that in an ME learning game $\mathcal{G}$, E's winning condition is simply to discover that $\hat{f}$ is interpretively blind, if he is. Call this condition *IB*. We establish the complexity of E's attempt to achieve *IB*. The first order case with a finite $\mathcal{H}$ where the game is restricted to moves m1,m2,m3, is rather trivial. More interesting is the case of an ME learning game $\mathcal{G} = (V^\infty, Win)$ with $Win = IB$ and in which E and $\hat{f}$ play higher order evaluation hypotheses.

Proposition 9. *Suppose an ME learning game $\mathcal{G} = (V^\infty, Win)$ with $Win = IB$ in which $\hat{f}$ plays moves described in (m4)- (m7). Then $\hat{f}$ is not interpretively blind iff play stops at some finite ordinal n.*

Suppose that in the play of $\mathcal{G}$, $\hat{f}$ accepts at some level n to add the sequence of evaluation hypotheses proposed by E. Then by the construction of the sequence and the requirement of coherence (constraint D), this confers upon some evaluation hypothesis $s{*}_1$ a non zero probability such that $P(T'|h^1_*) = 1$, where T' is incompatible with the body of evidence T. By accepting, $\hat{f}$ will have an evaluation hypothesis h^1_* with non zero probability such that $P(T'|h^1_*) = 1$, where T' is incompatible with the body of evidence T, which $\hat{f}$ has proposed as a source of learning (constraint E). Now when $\hat{f}$ updates his belief in T he must do so with respect to h^1_*, and he must now update his confidence in his evaluation hypotheses with respect not only to T but also T'. In that case, $P(h^1_*|T_n, T'_n) \not\twoheadrightarrow 0$ and $\mathrm{P}_n(T') \not\twoheadrightarrow 0$. As a result, $\hat{f}$ will be able to learn from T', and so he is not interpretively blind with respect to T.

If there is no stopping point at any finite ordinal, then E's is never able to get $\hat{f}$ to accept a T' positive hypothesis. In which case, $\hat{f}$ continues to only update on T and by Propositions 7 8, $\hat{f}$ is interpretively blind. □

Suppose E's winning condition for an ME learning $\mathcal{G}$, is to get $\hat{f}$ to accept a T' positive evaluation hypothesis. Call this winning condition for E $\mathcal{P}$ (for persuasion).

Corollary 3. *Suppose that in an ME learning game $\mathcal{G}$ with $Win = \mathcal{P}$. The complexity of Win is an R.E. set. If $Win = IB$ then Win is co-r.e. or Π_1 in the Borel Hierarchy.*

Proposition 10. *Suppose an ME learning game $\mathcal{G}$ with $Win = \mathcal{P}$ and $\hat{f}$ as described in Proposition 7. Then E has no winning strategy in $\mathcal{G}$.*

Proposition 8 implies $\hat{f}$'s evaluation hypotheses are updated on an argumentatively complete body of evidence T. When implemented via an ME game $\mathcal{G}$, the sequence of evaluation hypotheses in Proposition 7 provide a winning strategy for $\hat{f}$. Suppose E proposes an h^1 supporting e that is inconsistent with T. Even if E generates a suitable sequence of higher order T' positive evaluation hypotheses $h^1, h^2, h^3, \ldots$, given Constraint A above, $\hat{f}$ will only accept an evaluation hypothesis if he has no argument against it. But as T will eventually supply such an argument, $\hat{f}$ can always counter E's proposals. So she has no winning strategy. □

Not only is IB computationally complex (Corollary 3 shows it is not computable), Proposition 10 shows formally that even if E has rationally compelling arguments to show that $\hat{f}$ is better off (his payoff or reward is higher) in accepting her proposed sequence of evaluation hypotheses, $\hat{f}$ can rationally resort to T to counter her argument. Extracting someone from higher order IB is thus impossible by purely epistemic means. There is *no way* of getting someone, even a rational agent, out of higher order IB by purely epistemic arguments, given our assumptions. This pessimistic is borne out empirically: some people in the grip of right wing conspiracy theories in the US were dying of Covid19 in December of 2020 and January 2021 but continued to refuse to believe that it was that disease that was killing them—despite all the evidence and arguments they were given, they refused to let go of an obviously faulty but argumentatively complete T.

Of course, people sometimes *do* change their minds and do escape the grip of argumentatively complete theories, many times for epistemically exogenous reasons.[4] But by challenging one of our assumptions, rational agents can of course also reject IB. The weak link in our argument is assumption *A*, the "knowledge first" assumption. Perhaps $\hat{f}$ should accept evaluation hypotheses even if T attacks them. More likely, $\hat{f}$ should not accept all attacks equally. Probably, he should be skeptical of any body of evidence T that promotes PWMC for T and T eventually trustworthy evaluation hypotheses while attacking any point of view at variance with it.

We now explore the play between E and $\hat{f}$ in an ME learning game $\mathcal{G}$ where $Win = \mathcal{P}$ before $\hat{f}$ has accepted enough of the argumentatively complete T to close off learning from alternative bodies of evidence. Suppose T is argumentatively complete but comes in

[4] For instance, the satisfaction they derived from belonging to a particular community supported by a particular body of testimony might and does wane.

stages; if T'_i attacks T_i, then T_{i+1} but not T_i attacks T'_i. That is, an argumentatively complete T reacts to attacks but does not forsee all attacks in advance. Suppose a set of consistent first order evaluation hypotheses $\mathcal{H}^1 = \{h^1_1, h^1_2, ...\}$, with $P(h^1_1) = .6, P(h^1_2) = .4$, and $P(T_i|(h_1) = 1 = P(T'_i|h_2)$. Now suppose $T'_1 \cup T_1$ is inconsistent and E proposes h^1_2 since $h^1_2 \models T'_i$. Since the h^1_i are consistent, $P(T_1|h^1_2) = 0 = P(T'_2|h^1_1)$. At this point, $\hat{f}$ could accept E's proposal under constraint (A), $\mathcal{G}$ ends and E wins. $\hat{f}$ will continue to update over stages T and T' with the marginal probabilities $P(T_i) = .6$ and $P(T'_i) = .4$ remaining stationary.

On the other hand, $\hat{f}$ may decide to wait to see what the next stage T_2 of T brings. As T is argumentatively complete, T_2 will attack T'_2, and add a nullifying $h^2 \in \mathcal{H}^2$ supported by T_2. Should $\hat{f}$ accept h^2, the probability of h^1_2 will go to 0 in $\mathcal{H}$. But now suppose we have a constraint, Discount, that discounts any nullifying sequence from T. It would be unreasonable for $\hat{f}$ to wipe out alternatives in the face of this level of uncertainty; at this stage, $P(T_2) = .6$ and $P(T'_2) = .4$. Summarizing:

Proposition 11. *Suppose an ME learning game $\mathcal{G}$ with constraint A replaced by Discount and with Win = $\mathcal{P}$ and $\hat{f}$ as described in Proposition 7. E then has a winning strategy in G, and IB does not arise for $\hat{f}$.*

7 CONCLUSIONS

Interpretive blindness results from a dynamic, iterative process whereby a learner's background beliefs and biases lead her to update her beliefs based on a body of testimony T, and then biases inherent in T come back to reinforce her beliefs and her trust in T's source(s), further biasing her towards these sources for future updates. We have introduced and formally characterized IB. We have shown that IB can prevent learning even in higher order Bayesian frameworks for learning from argumentatively complete testimony, despite the presence of constraints designed to promote good epistemic practices. We also shown that IB is computationally complex as a co-r.e. set via a game theoretic analysis, and that an agent may rationally remain in IB in the face of epistemic arguments. Our game theoretic analysis can also be extended to cases where the agent falls out of IB but then is a recidivist and becomse a prisoner once more. We leave that for future work.

How general are the results in Propositions 7 and 8? PAC, Statistical Physics Framework, VC, and supervised Bayesian learning are four different instantiations of Wolpert's extended Bayesian formalism that we use [Wolpert 2018]. Thus our results should hold for other frameworks.

Investigating IB alas is not just an academic enterprise. IB really does happen, with sometimes tragic or dangerous results. We think a careful formal analysis is urgent for society. Finally, we note that while we have focused on IB as a problem for learning from testimony, the problem it raises for learning extends to any case in which we do not have unmediated access to ground truth and our data is "theory laden" [Hanson 1958].

ACKNOWLEDGMENTS

This work was supported by the AI Interdisciplinary Institute ANITI, funded by the French program "Investing for the Future – PIA3" under Grant agreement no. ANR-19-PI3A-0004, the ANR PRCI grant SLANT (no. ANR-19-CE23-0022), and also by the ICT 38 H2020 grant COALA (no. 957296).

REFERENCES

Leila Amgoud and Robert Demolombe. 2014. An argumentation-based approach for reasoning about trust in information sources. *Argument and Computation* 5:2-3 (2014), 191–215.

Nicholas Asher and Julie Hunter. 2021. Interpretive blindness and the impossibility of learning from testimony. In *Proceedings of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2021)*, U. Endriss, A. Nowé, F. Dignum, and A. Lomuscio (Eds.). Extended abstract.

Nicholas Asher and Soumya Paul. 2018. Strategic conversation under imperfect information: epistemic Message Exchange games. *Logic, Language and Information* 27.4 (2018), 343–385.

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. 2002. Finite-time analysis of the multiarmed bandit problem. *Machine learning* 47, 2-3 (2002), 235–256.

Jeffrey S Banks and Rangarajan K Sundaram. 1994. Switching costs and the Gittins index. *Econometrica: Journal of the Econometric Society* (1994), 687–694.

Apostolos N Burnetas and Michael N Katehakis. 1997. Optimal adaptive policies for Markov decision processes. *Mathematics of Operations Research* 22, 1 (1997), 222–255.

Quassim Cassam. 2016. Vice epistemology. *The Monist* 99, 2 (2016), 159–180.

Christiano Castelfranchi and Rino Falcone. 2010. *Trust theory: A socio-cognitive and computational model*. Vol. 18. John Wiley & Sons.

Nicolo Cesa-Bianchi and Gábor Lugosi. 2006. *Prediction, learning, and games*. Cambridge university press.

Benoit Dardenne and Jacques-Philippe Leyens. 1995. Confirmation Bias as a Social Skill. *Personality and Social Psychology Bulletin* 21.11 (1995), 1229–1239.

Igor Douven and Christoph Kelp. 2013. Proper bootstrapping. *Synthese* 190, 1 (2013), 171–185.

Phan Minh Dung. 1995. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial intelligence* 77, 2 (1995), 321–357.

Aurélien Garivier and Olivier Cappé. 2011. The KL-UCB Algorithm for Bounded Stochastic Bandits and Beyond.. In *COLT*. 359–376.

Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. 2013. *Bayesian data analysis*. CRC press.

Norwood Russell Hanson. 1958. *Patterns of discovery: An inquiry into the conceptual foundations of science*. Vol. 251. Cambridge University Press.

Anna Ichino and Juha Räikkä. 2020. Non-doxastic conspiracy theories. *Argumenta* 2020 (2020), 1–18.

Kathleen Hall Jamieson and Joseph N Cappella. 2008. *Echo chamber: Rush Limbaugh and the conservative media establishment*. Oxford University Press.

Kenji Kawaguchi, Leslie Pack Kaelbling, and Yoshua Bengio. 2017. Generalization in deep learning. *arXiv preprint arXiv:1710.05468* (2017).

Thomas Kelly. 2008. Disagreement, dogmatism, and belief polarization. *The Journal of philosophy* 105, 10 (2008), 611–633.

Thomas L Griffiths, Charles Kemp, and Joshua B Tenenbaum. 2008. Bayesian models of cognition. (2008).

Tze Leung Lai and Herbert Robbins. 1985. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics* 6, 1 (1985), 4–22.

Jouko Lampinen and Aki Vehtari. 2001. Bayesian approach for neural networks—review and case studies. *Neural networks* 14, 3 (2001), 257–274.

Charles G. Lord, Lee Ross, and Mark R. Lepper. 1979. Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology* 37.11 (1979), 2098–3009.

Elijah Millgram. 2015. *The great endarkenment: philosophy for an age of hyperspecialization*. Oxford University Press.

Kevin Murphy, Yair Weiss, and Michael I Jordan. 2013. Loopy belief propagation for approximate inference: An empirical study. *arXiv preprint arXiv:1301.6725* (2013).

Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nati Srebro. 2017. Exploring generalization in deep learning. In *Advances in Neural Information Processing Systems*. 5947–5956.

C Thi Nguyen. 2020. Echo chambers and epistemic bubbles. *Episteme* 17, 2 (2020), 141–161.

Raymond S Nickerson. 1998. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology* 2.2 (1998), 175–220.

Margit E. Oswald and Stefan Grosjean. 2004. Confirmation bias. In *Cognitive illusions: A handbook on fallacies and biases in thinking, judgement and memory*, Rüdiger F. Pohl (Ed.). Hove, UK: Pyschology Press, 79–96.

Joshua B Tenenbaum, Thomas L Griffiths, and Charles Kemp. 2006. Theory-based Bayesian models of inductive learning and reasoning. *Trends in cognitive sciences* 10, 7 (2006), 309–318.

Joshua B Tenenbaum, Charles Kemp, Thomas L Griffiths, and Noah D Goodman. 2011. How to grow a mind: Statistics, structure, and abstraction. *science* 331, 6022 (2011), 1279–1285.

Amos Tversky and Daniel Kahneman. 1975. Judgment under uncertainty: Heuristics and biases. In *Utility, probability, and human decision making*. Springer, 141–162.

Amos Tversky and Daniel Kahneman. 1985. The framing of decisions and the psychology of choice. In *Environmental Impact Assessment, Technology Assessment, and Risk Analysis*. Springer, 107–129.

Jonathan Vogel. 2008. Epistemic bootstrapping. *The Journal of Philosophy* 105, 9 (2008), 518–539.

Jonathan Weisberg. 2010. Bootstrapping in general. *Philosophy and Phenomenological Research* 81, 3 (2010), 525–548.

Peter Whittle. 1980. Multi-armed bandits and the Gittins index. *Journal of the Royal Statistical Society. Series B (Methodological)* (1980), 143–149.

Timothy Williamson. 2002. *Knowledge and its Limits*. Oxford University Press.

David H Wolpert. 2018. The relationship between PAC, the statistical physics framework, the Bayesian framework, and the VC framework. In *The mathematics of generalization*. CRC Press, 117–214.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 2016. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530* (2016).