Interpretive Blindness and the Impossibility of Learning from Testimony

Extended Abstract

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

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

ABSTRACT

We model *interpretive blindness*, a type of epistemic bias that poses a problem for learning from testimony, in which one acquires information from text or conversation but lacks direct access to ground truth. Interpretive blindness arises when a co-dependence between background beliefs and interpretation leads to a dynamic process of bias hardening that impedes or precludes learning. We argue that when bodies of data are *argumentatively complete*, even constraints from hierarchical Bayesian learning designed to promote good epistemic practices will fail to stop interpretive blindness.

KEYWORDS

Learning, bias, agent modeling, Bayesian learning

ACM Reference Format:
Nicholas Asher and Julie Hunter. 2021. Interpretive Blindness and the Impossibility of Learning from Testimony: Extended Abstract. In *Proc. of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2021), Online, May 3–7, 2021*, IFAAMAS, 3 pages.

1 INTRODUCTION

In this paper, we introduce and analyze *interpretive blindness*, a type of epistemic bias exemplified by humans that has not yet, as far as we know, been theoretically examined. Interpretive blindness is a special problem for learning from testimony, in which one acquires information or data about some phenomenon P from, say, books, news outlets, social media or conversation, without having direct access to P. Learning from testimony requires not only trusting the source that conveys and promotes the testimony but also being biased towards some sources more than others, so that a learner can make a decision in the face of conflicting bodies of testimony. Restricting one's attention to a limited set of sources, however, leads all too easily to the hardening of one's biases towards those sources in such a way that a learner becomes effectively blind to bodies of testimony that are incompatible with or not entailed by the bodies of testimony promoted by one's favored sources.

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. It is related to the framing biases of [14, 15] and to confirmation bias [5, 10, 12, 13], in which agents interpret new evidence in a way that confirms their beliefs. These forms of bias, however, concern how beliefs and bias influence interpretation, painting only part of the picture of interpretive bias (see also [3]). Our interest here is in how performing a Bayesian update of one's beliefs based on a given interpretation of a body of data can engender bias hardening and preclude learning: when confronted with evidence that contradicts their beliefs, interpretively blind agents will discount it outright, no matter how reasonable or well-founded it might be.

Proc. of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2021), U. Endriss, A. Nowé, F. Dignum, A. Lomuscio (eds.), May 3–7, 2021, Online.

2 DEFINING INTERPRETIVE BLINDNESS

A body of testimony T is a collection of information conveyed by a given source s such as *The New York Times*, *Fox News*, *Facebook*, *4Chan*, or a particular individual or set of individuals. Such bodies of information are *dynamic* in that they evolve over time as they are updated with new facts and events. In other words, T comes in "stages", so that $T = \{T_1, T_2, ..., T_n, ...\}$, where stages might be defined by times or even conversational turns, and each stage T_i is the body of evidence accumulated up to stage i.

Learning from a body of testimony T with source s requires a learner $\hat{f}$ to judge T as credible, a judgment that will depend on s's evaluation of T (whether s promotes or challenges T), as well as $\hat{f}$'s antecedent hypotheses about s. Let $\mathcal{H}$ be a set of *evaluation hypotheses*, where each $\mathfrak{h} \in \mathcal{H}$ gives the evaluation of a set $\mathcal{T}$ of bodies of testimony T relative to a source s. $\mathfrak{h} \in \mathcal{H}$ defines a conditional probability $P(T|\mathfrak{h})$ for $T \in \mathcal{T}$, which we will sometimes write as $\mathfrak{h}(T)$, where $\mathfrak{h}(T) = 0$ means T is untrustworthy according to $\mathfrak{h}$, and $\mathfrak{h}(T) = 1$ means T is trustworthy (s fully endorses T). Following Wolpert's 2018 extended Bayesian framework, our learner $\hat{f}$ updates his belief in T relative to $\mathcal{H}$.

Our learner $\hat{f}$ will have a probability distribution over his evaluation hypotheses $\mathcal{H}$. Given the co-dependence of beliefs and evidence, this distribution is updated relative to the stages of T as it develops. This is intuitive; the testimony T should serve as evidence upon which $\hat{f}$ updates his beliefs, including his judgment about the source of T. But the co-dependence tells us that $\hat{f}$ updates his confidence in T via these updated beliefs. Let $\mathsf{E}_n(\mathfrak{h}_i)$ be the expected value of $\mathfrak{h}_i$ after conditionalizing on T_n, i.e. $P(\mathfrak{h}_i|T_n)$ and $\mathsf{E}_n(T)$ the expected marginal value of T after n updates—i.e., $\Sigma_{\mathfrak{h}\in\mathcal{H}} P(T_n|\mathfrak{h})$.

Proposition 1. Suppose $T = \{T_1, T_2, ..., T_n, ...\}$ is a dynamic body of evidence, with a set of hypotheses $\mathcal{H} = \{\mathfrak{h}_1, \mathfrak{h}_2, ...\mathfrak{h}_k : T \rightarrow [0, 1]$ for $T \in \mathcal{T}\}$ with $P(T_i|\mathfrak{h}_1) = 1$ and $\mathfrak{h}_1$ with non zero probability. Under certain mild assumptions about $\mathfrak{h}_k \in \mathcal{H}$ and the probability distribution over $\mathcal{H}$, for $T \neq T'$, iterated updating of

probabilities over $\mathcal{H}$ based on T_i yields:

$$\text{As } n \to \infty,\ \mathsf{E}_n(T') \to 0 \text{ and } \mathsf{E}_n(T) \to 1.$$

Proposition 1 captures interpretive blindness: through iterative updating given the codependence of beliefs and interpretation, $\hat{f}$ has put all of its subjective probability mass on a set of evaluation hypotheses that count only some bodies of evidence trustworthy.

Interpretive blindness precludes learning from any other body of evidence that is not promoted by one's favored sources. To learn a hypothesis h, $\hat{f}$'s estimation of h at some stage should be closer to the objective or ideal assignment (posterior) h_p to h, than her prior probability for h. Similarly for marginal probabilities: $E_n(x)$ should track x_p, the posterior of x, given a random sampling of data $x \in X$. We consider loss functions $\mathcal{L}(E_n(h), h_p)$ and $\mathcal{L}(E_n(x), x_p)$. 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 h if additional evidence does not eventually decrease loss; i.e. we cannot show $lim_{n\to\infty}\mathcal{L}(\mathsf{E}_n(h), h_p) < \mathcal{L}(E_0(h), h_p)$.

Proposition 2. Let $\hat{f}$ be a Bayesian learner with source functions and bodies of evidence T, T' as in Proposition 1 and suppose all evidence e confirming a hypothesis h is such that $T' \models e$. Then $\hat{f}$ is incapable of learning h.

3 HIERARCHICAL BAYESIAN LEARNING

Being interpretively blind to a body of testimony T' might not be a problem if T' is one that $\hat{f}$ *should* discount (because, e.g., it contains objective falsehoods). But to make sure $\hat{f}$ does not discount relevant evidence, we need to add constraints on $\hat{f}$'s beliefs. Hierarchical Bayesian models were designed to address this problem [7]. In these models, a first order Bayesian learning model like the one we have discussed in Section 2 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. Factors like the consistency or the predictive accuracy of a testimony source might be important, or the extent to which testimony from other sources agrees with its content. A third level could then involve constraints on those constraints or arguments for the second level constraints.

Simply requiring evaluation hypotheses that obey exogenous constraints, however, begs the question of why $\hat{f}$ should accept them. A body of testimony T can be *argumentatively complete*, meaning that it can explicitly respond to and argue with any doubts raised by data in conflict with T that might attack T's credibility. A skillful climate denier, for example, will always find a way to undercut the most scientifically careful argument, if only by attacking the reliability of the source of the argument. Argumentatively complete testimony makes learning—and, as a result, teaching and persuading—impossible in certain cases. We formally develop this notion in [2] to define an argumentatively complete body of testimony and to prove our main result, which we provide here:

Proposition 3. Suppose T is argumentatively complete. Let $\hat{f}$ be a Higher Order Bayesian learner whose evaluation hypotheses: are coherent, make T potentially trustworthy and are updated on T. If for $T' \neq T$, T' confirms a hypothesis $\mathfrak{h}$ that T does not, then $\hat{f}$ is incapable of learning $\mathfrak{h}$.

It is clear what has gone wrong: $\hat{f}$ should impose constraints on the evaluation hypotheses $\mathcal{H}$ that would minimize the losses $\mathcal{L}(\mathsf{E}_n(h), h_p)$ and $\mathcal{L}(E_n(x), x_p)$ of Proposition 2, but with belief based only on testimony, $\hat{f}$ will not have access to h_p or to x_p. In the face of testimony T' that contradicts T, $\hat{f}$ should, as a good Bayesian, conditionalize on T', yet given that T and T' rely on sources, it's not obvious which should be used to revise one's beliefs: the familiar body of evidence or evidence posed by another source that might not be trustworthy. Ideally, $\hat{f}$ would investigate the inconsistencies in $T \cup T'$ and find other evidence that confirms or disconfirms T or T'. But that might not be possible, and in any case, T provides ready-made arguments for rejecting any T' distinct from T.

Proposition 3 should generalize to other frameworks: [16] argues that PAC, Statistical Physics Framework, VC, and supervised Bayesian learning are different instantiations of an extended Bayesian formalism, which is a slight extension of our framework.

[2] gives the formal details of the concepts presented here and develops a game theoretic framework to investigate the complexity of interpretive blindness, with suggestions for how to escape it.

4 COMPARISONS TO PRIOR WORK

Interpretive blindness is an epistemological bias, and in particular, a kind of iterated confirmation bias [10, 12, 13] brought on by the natural co-dependence of beliefs and interpretation of evidence. Unlike much of the psychological literature which finds epistemologically exogenous justifications for this bias [5], however, we have shown how interpretive blindness is a natural outcome of Bayesian updating, rational resource management and the belief interpretation co-dependence. As far as we know, this phenomenon has not been studied with rigorous techniques beforehand.

Interpretive blindness is also related to work on generalization. Epistemic biases affect generalization and learning capacity in ways that are still not fully understood [8, 9, 11, 17]. Work on argumentation [1, 6] is also relevant to interpretive blindness, and there are important connections to the literature on trust [4]; in our set up, learning agents trust certain sources over others, and our higher order setting invokes a hierarchy of reasons. Nevertheless, 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, we have argued that interpretive blindness results from the dynamic nature of the Bayesian framework, with beliefs evolving under changing evidence. It is this dynamic evolution that is crucial to our main points 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.

ACKNOWLEDGMENTS

We gratefully acknowledge the support of the project SLANT (ANR-19-CE23-0022) and the AI Institute ANITI (ANR-19-PI3A-0004).

REFERENCES

[1] 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.

[2] Nicholas Asher and Julie Hunter. 2021. Interpretive blindness: a challenge for learning from testimony. (2021). unpublished manuscript.

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

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

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

[6] 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.

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

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

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

[10] 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.

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

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

[13] 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.

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

[15] 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.

[16] 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.

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