Interpretive Blindness and the Impossibility of Learning from Testimony

Nicholas Asher† and Julie Hunter‡

†CNRS-IRIT, Toulouse and ‡ LINAGORA Labs, Toulouse

We model interpretive blindness (IB), 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 for a Bayesian learner $\hat{f}$.

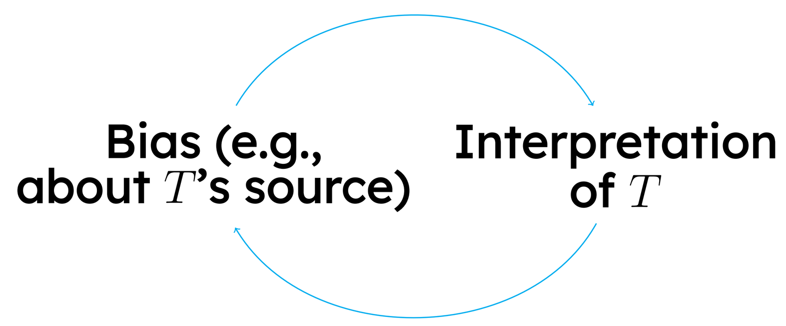
Bias: A Double-Edged Sword

Learning from testimony T requires evaluating T's (or the source of T's) reliability.

But restriction to a limited set of sources S can lead to the hardening of biases towards them and a blindness to bodies of testimony incompatible with or not entailed by those promoted by S.

$\Rightarrow$ a dynamic, iterative process

Co-Dependence of Beliefs and Interpretation (CoBI)



Bodies of Testimony

A body of testimony T: a collection of information conveyed by a source s (*The New York Times*, an individual...)

Such bodies T are dynamic: T comes in cumulative "stages", $T = \{T_1, T_2, ..., T_n\}$, delimited by conversational turns, times, etc.

Evaluation Hypotheses

A set of evaluation hypotheses $\mathcal{H}$: each $h \in \mathcal{H}$ evaluates a set $\mathcal{T}$ of bodies of testimony T relative to a source s.

$h \in \mathcal{H}$ defines a conditional probability $P(T|h)$ for $T \in \mathcal{T}$

- $h(T) = 0$ when T is untrustworthy according to h
- $h(T) = 1$ when T is trustworthy, i.e., h fully endorses T

$\hat{f}$ updates his belief in T relative to $\mathcal{H}$ (Wolpert, 2018).

Interpretive Blindness

CoBI tells us that $\hat{f}$ will put all subjective probability mass on a set $\mathcal{H}$ that counts only some T as trustworthy.

Let $P_{\mathcal{H}}$ be $\hat{f}$'s probability distribution over $\mathcal{H}$: $P_{\mathcal{H}}$ is updated iteratively as T develops.

E$_n(h_i)$: expected value of h_i after conditionalizing on T_n, i.e. $P(h_i|T_n)$

But CoBI tells us that $\hat{f}$ updates his confidence in T via these updated beliefs.

E$_n(T)$: expected value of T after n updates, $P(T_n|h)$

Proposition 1: For $T = \{T_1, T_2, ..., T_n, ...\}$ and $\mathcal{H} = \{h_1, h_2, ...h_k\}$, suppose $P(T_i|h_1) = 1$, $P(T_i|h_j) < .5, j \neq 1$ and h_1 has non-0 probability. For $T \not\models T'$ and $T' \not\models T$, iterated updating of probabilities over $\mathcal{H}$ based on T_i yields:

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

Learning

IB precludes learning from 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 assignment (posterior) h_p to h, than her prior probability for h.

$\hat{f}$ cannot learn h if additional evidence does not eventually decrease loss; i.e. we cannot show $lim_{n \to \infty} \mathcal{L}(\text{E}_n(h), h_p) < \mathcal{L}(E_0(h), h_p)$.

Proposition 2: Let T, T', and $\hat{f}$'s source functions, $\mathcal{H}$, be 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.

Hierarchical Bayesianism

Hierarchical Bayesian models add constraints on beliefs to ensure that a learner $\hat{f}$ does not discount relevant evidence (Gelman et al., 2013).

Level 1: a first order Bayesian learning model with certain parameters, e.g., our evaluation hypotheses h.

Level 2: a Bayesian learning model detailing factors allowing for a reliable estimation of a hypothesis h's accuracy.

- internal consistency, consistency with other sources, predictive accuracy, ...

Level 3: constraints on, or arguments for, Level 2 constraints.

...and so on.

Argumentative Completeness

But if we try to require hypotheses h that obey exogenous constraints, why should our higher-order learner $\hat{f}$ accept them?

An argumentatively complete (AC) T: explicitly responds to and argues with any doubts raised by data in conflict with T.

AC testimony can make learning impossible in a higher order setting.

Proposition 3: Let T be AC and suppose $\hat{f}$'s evaluation hypotheses: are coherent, make T potentially trustworthy and are updated on T. If for $T' \neq T$, T' confirms a hypothesis h and T does not, then $\hat{f}$ is incapable of learning h.

- $\hat{f}$ cannot impose constraints on $\mathcal{H}$ to minimize $\mathcal{L}(\text{E}_n(h), h_p)$, as $\hat{f}$ has no access to h_p
- $\hat{f}$ should conditionalize on T', but T''s source might be untrustworthy
- $\hat{f}$ should investigate inconsistencies in $T \cup T'$, but T provides ready-made arguments for rejecting T'

See Asher & Hunter (2021) for more.

Related Concepts

Confirmation bias concerns how beliefs and bias influence interpretation.

- we look at how, given a certain interpretation of evidence, Bayesian update on one's beliefs can engender bias hardening and preclude learning
- IB agents will discount even reasonable, well-founded evidence laid directly before them if it contradicts their beliefs

Work on **argumentation** and **trust** tends to consider *static* constraints one can impose on inference in the face of a possibly inconsistent belief base.

- IB results from the *dynamic* nature of the Bayesian framework, with beliefs evolving under changing evidence
- we are not looking at the problem of consistency, but rather the problems of *entrenchment* and bias

References for Abstract

Amgoud & Demolombe, 2014. An argumentation-based approach for reasoning about trust in information sources; **Asher & Hunter, 2021.** Interpretive blindness: a challenge for learning from testimony; **Asher & Paul, 2018.** Strategic conversation under imperfect information: epistemic Message Exchange games; **Castelfranchi & Falcone, 2010.** Trust theory: A socio-cognitive and computational model; **Dardenne & Leyens, 1995.** Confirmation Bias as a Social Skill; **Dung, 1995.** On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games; **Gelman et al., 2013.** Bayesian data analysis; **Kawaguchi et al., 2017.** Generalization in deep learning; **Lampinen & Vehtari, 2001.** Bayesian approach for neural networks—review and case studies; **Lord et al., 1979.** Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence; **Neyshabur et al., 2017.** Exploring generalization in deep learning; **Nickerson, 1998.** Confirmation bias: A ubiquitous phenomenon in many guises; **Oswald & Grosjean, 2004.** Confirmation bias; **Tversky & Kahneman, 1975.** Judgment under uncertainty: Heuristics and biases; **Tversky & Kahneman, 1985.** The framing of decisions and the psychology of choice; **Wolpert, 2018.** The relationship between PAC, the statistical physics framework, the Bayesian framework, and the VC framework; **Zhang et al., 2016.** Understanding deep learning requires rethinking generalization.