What is bias?

Kevin Dorst 24.223 Rationality

I. Statistical bias

Putative cases of confirmation bias:

- Wason, number-progression rules. "Positive test strategy."
- Pseudodiagnositicity. *Hypothesis:* Jim is an introvert.
- Biased assimilation.
- Selective exposure.

2-4-6 satisfies rule. What is rule?

"Do you ever like to be alone?"

Kelly 2008

Check NYT or WSJ?

Every inductive or decision method *sometimes* misfires. If we know the details of how it works, we can even *predict* when it misfires. So how can we assess whether the deviation is irrational?

"Believe in accord with your evidence" misfires whenever your evidence is misleading...

Proposal: bias as *expected* deviation from accurate belief / best decision. Irrational bias as expected deviation that is *common* and *costly*.

Why not say just any expected deviation?
H&H: because even *optimal* (Bayesian) beliefs/decisions sometime exhibit expected deviations from accurate beliefs.

Let e_X be an estimator of X, i.e. a function from data/evidence to numbers that are your best estimate of a variable X.

e_X is a *statistically unbiased estimator* of X[1] iff for all thresholds t, $\mathbb{E}_P(e_X|X = t) = t$.

[1] wrt P!

$\Leftrightarrow \forall t : \mathbb{E}_P(e_X - X|X = t) = 0$.

Wrt which distribution? H&H don't say, presumably because they think it won't matter. Either subjective or objective probabilities will (on their definition) often agree.

We'll come back to this...

Fact: so-defined, Bayesian estimators are biased. More generally, there is a **bias-variance tradeoff**.

Lower-variance estimators are less misled by misleading data (less overfitting), but exhibit more bias. Unbiased estimators have high variance and are prone to overfitting.

Example: X = *the bias of this coin*. We'll flip it 10 times.

- Unbiased estimator: proportion heads. ("Frequentist estimator")
  But high variance—likely to be inaccurate.
- Biased estimator: mean of Bayesian posterior that begins uniform over biases.
  Biased: conditional on $X = 1$, expected estimate is $\mathbb{E}_P(e_X|\mathit{10\ heads}) = \mathit{Mean}(\mathit{Beta}(11, 1)) = \frac{11}{12} \approx 0.92 < 1 = X$.

Clear when toss only 1 or 2 times.

Beta(1,1) prior. If see k heads and $10 - k$ tails, go to $\mathit{Beta}(1 + k, 1 + 10 - k)$

Fact: Expected[2] accuracy of Bayesian posterior is higher than that of proportion-heads.

[2] Relative to Bayesian priors! Or objective ones if we sample from coin biases uniformly.

So, they conclude, bias can be good!

II. Bayesian bias?

Is this the right definition of bias?

e_X is a *Bayesian-unbiased estimator* of X iff $\mathbb{E}_P(e_X - X) = 0$.

Iff $\mathbb{E}_P(e_X) = \mathbb{E}_P(X)$

On this definition, there need be no bias-variance tradeoff. The above Bayesian posterior is unbiased!

How do the two definitions do across cases?

1) Your future estimate of X, after learning it's value.

2) Your posterior estimate of X, after conditioning on the true cell of a partition.
Suppose the partition is trivial, so you don't learn anything: $\Pi = \{W\}$. Their definition says your posterior is biased!

E.g. an indicator about X

3) Conglomerability failures are biased.
Bill is delusional, so that no matter what he sees, he'll increase his confidence that it landed heads, e_X, to 0.8.

Both definitions agree

$$\begin{aligned}\mathbb{E}(e_X - X) &= P(X=1)(0.8-1) + P(X=0)(0.8-0)\\ &= 0.5 * (-0.2) + 0.5 * 0.8 = 0.3\end{aligned}$$

Is bias necessarily bad? No:

A biased *but useful* estimate: All Jill knows is that I'll flip a fair coin. But you and I know that if it lands heads ($X = 1$), I'll tell her it did, and if it lands tails ($X = 0$) I'll tell her nothing.

e_X = Jill's future credence: if $X = 1$, then $e_X = 1$; and if $X = 0$, then $e_X = 0.5$. So biased:

$$\begin{aligned}\mathbb{E}(e_X - X) &= P(X=1)(1-1) + P(X=0)(0.5-0)\\ &= 0.5 * 0 + 0.5 * 0.5 = 0.25\end{aligned}$$

Your prior expects Jill's future credence in heads to be more accurate than your own.

Pros and cons of these alternative definitions of bias?