Hedden 2019: Hindsight bias

Kevin Dorst 24.223, Rationality

I. Hindsight bias

> **Railroad.** You are presented with a railroad company's evidence about a potentially-dangerous piece of railroad track. This includes expert evaluations, the company's assessment, and a warning from the local authorities. You weigh these up and judge that the probability of a derailment is 40%. Then you're told that a train *in fact* derailed. Thinking back to your earlier ("*ex ante*") evidence, you now think that it supported a derailment to degree 70%, and are more likely to find the railroad company negligent.

Other cases: potentially-violent patient; earlier-self's predictions.

Why think this is irrational?

1) Evidence can be misleading, so there's no inconsistency between H being (ex ante) unlikely and nevertheless happening.
2) The (later) truth of H can't affect what the (earlier) evidence for it was! The judgment is about what the *evidence itself* supported.

Hedden claims that even though (1) and (2) are both true, rational Bayesians will nonetheless exhibit hindsight bias.

II. What is hindsight bias?

Let E be the (ex ante) body of evidence, and let $S_E(H)$ be the degree to which E supports H.

$\approx$ the credence an ideally rational agent with total evidence E would have in H.

Let cr be your credence function. (This is 'P', in Hedden's notation.)

Two different formalizations of hindsight bias:

1) *Threshold-raising HB*. Fix a threshold t, and say that E supports H iff $S_E(H) > t$.

   Firmness

   You exhibit threshold-raising hindsight bias iff learning H raises your credence that E supports H:

$$cr(S_E(H) > t \big| H) \;>\; cr(S_E(H) > t)$$

Example:
$cr(S_E(H) > t) = 0.4$, and yet
$cr(S_E(H) > t|H) = 0.5$.

2) *Estimate-raising HB*. You exhibit estimate-raising hindsight bias iff learning H raises your expectation of $S_E(H)$:

   $\approx$ probabilifying

$$\mathbb{E}_{cr}(S_E(H)\big|H) > \mathbb{E}_{cr}(S_E(H))$$

$$\text{i.e.} \quad \sum_{x_i \in \mathbb{R}} cr(S_E(H) = x_i|H) \cdot x_i \;>\; \sum_{x_i \in \mathbb{R}} cr(S_E(H) = x_i) \cdot x_i$$

Example:
$\mathbb{E}_{cr}(S_E(H)) = 0.6$, and yet
$\mathbb{E}_{cr}(S_E(H)|H) = 0.625$.

III. Why hindsight bias can be rational

First notice theres nothing necessarily non-Bayesian about hindsight bias. Here's a simple model:

Let the threshold for "support" be $t = 0.7$. Suppose you know that either $S_E(H) = 0.5$ or $S_E(H) = 0.75$, and are 60-40 split between them:

	H	$\neg H$
$S_E(H) = 0.75$	0.3	0.1
$S_E(H) = 0.5$	0.3	0.3

Then *cr* exhibits threshold-raising HB:
$cr(S_E(H) > 0.7) = 0.4$, yet
$cr(S_E(H) > 0.7|H) = 0.5$.

Moreover *cr* exhibits estimate-raising HB:
$\mathbb{E}_{cr}(S_E(H)) = 0.6 \cdot 0.5 + 0.4 \cdot 0.75 = 0.6$, while
$\mathbb{E}_{cr}(S_E(H)|H) = 0.5 \cdot 0.5 + 0.5 \cdot 0.75 = 0.625$.

Hedden further argues that rational Bayesians not only *can* exhibit hindsight bias, but they *very often will*.

This follows from two claims:

i) *cr* will often be uncertain what $S_E(H)$ is.

$cr(S_E(H) = n) < 1$, for all n.

ii) *cr* should think that $S_E(H)$ is correlated with the truth of H.

For all n, $cr(H|S_E(H) > n) > cr(H)$.

Why? Relevance is symmetric! $cr(A|B) > cr(A)$ iff $cr(B|A) > cr(B)$.

Let $A = H$ and let $B = [S_E(H) > t]$.
(ii) implies that $cr(H|S_E(H) > n) > cr(H)$. So it follows that
$cr(S_E(H) > n|H) > cr(S_E(H) > n)$. That's threshold-raising HB!

Since (ii) applies to *all* thresholds n, also works for estimate-raising HB.

Why accept (i)?

E is your evidence at t_1—the expert said blah, the company said bleh...

So your credence is $cr(H) = cr_0(H|E)$.

cr_0 your hypothetical prior.

The *ideal* support function is $S_E(H) = S_0(H|E)$.

S_0 is the *ideal* prior, whatever it is.

You may be rational and yet be unsure what the ideal posterior is: $cr(S_E(H) = 0.75) > 0$ and $cr(S_E(H) = 0.5) > 0$.

In fact, you *should* be, says Hedden, because you should be unsure how to trade off the theoretical virtues of various competing hypotheses.

Why accept (ii)?

cr should think that $S_E(H)$ is a guide to truth—otherwise, it wouldn't be the *ideal* credence.