Hahn and Harris 2014: What Does it Mean to be Biased?

Kevin Dorst
kevindorst@pitt.edu

Rationality Seminar
March 8, 2021

I. "Bias" in Psychology: Historical Overview

Confirmation bias first widely-discussed bias. "Tendency to seek an interpret evidence in a way that favors current hypothesis."

Wason (1960) number-progression rules. E.g. said '2-4-6' satisfies rule; have to query experimenter to try to suss out rule.

Some (not majority) fail to realize rule is 'increasing numbers', since they cotton on to 'increasing evens'.

Search strategy seems to contradict Popperian "try to falsify". But subtle! Suppose $P(\textit{incr. evens}) = 0.7$. Then asking '4-6-8?' may well falsify!

"Seeking confirmation" doesn't mean you're looking for evidence you *expect* to raise your credence. If you're asking an unambiguous question and will update on the answer, you *can't* expect that!

> Let Q be a question $\{q, \neg q\}$, P^+ your future credence. If q, your future credence in p is $P^+(p) = P(p|q)$; if $\neg q$, your future credence is $P^+(p) = P(p|\neg q)$.
>
> So what's your expectation of your future credence, $\mathbb{E}(P^+(q))$? By total expectation:
>
> $$\begin{aligned}\mathbb{E}(P^+(q)) &= P(q) \cdot \mathbb{E}(P^+(q)|q) + P(\neg q) \cdot \mathbb{E}(P^+(q)|\neg q)\\ &= P(q) \cdot P(p|q) + P(\neg q) \cdot P(p|\neg q)\end{aligned}$$
>
> Which, by total probability, equals your current credence $P(p)$.

Reasoning generalizes: if you're going to update on a question (partition) by conditioning, your expectation for your future credence equals your current credence (Weisberg 2007; Briggs 2009).

Q: So what does "seeking confirmation" *mean* in a context like this, where you're asking unambiguous questions?

Choose option with "best chance of confirming"? Holding fixed $P(q)$, choosing a q for which $P(p|q) > P(p)$ and maximizes $P(q)$ will be such that $P(p|\neg q) = 0$ and $P(p|q) - P(p)$ is small:

$$P(p) = P(q) \cdot P(p|q) + P(\neg q) \cdot P(p|\neg q)$$

Klayman and Ha (1987): "positive test strategy"? $\approx$ examining instances which, if current hypothesis is true, will fall under it's scope. (I.e. answer 'yes' if hypothesis true.)

Q: What does this look like in generality. Is it an issue? If someone wants to **write a paper** on this, please do...

Back to Hahn and Harris (2014): "confirmation bias" has become a bit of a catchall. Sometimes about how/where one searches for information; sometimes about how one interprets it or remembers it; sometimes about being overconfident; sometimes about giving more weight to initial batch of evidence; etc.

Next bias: **conservatism**

Method: draw colored chips from mystery bag; ask them how likely it is to be bag 1 vs. bag 2. E.g. B1 = 5 red, 5 black; B2 = 7 red, 3 black. First two draws red. How likely to be B2?

$$\begin{aligned} P(B_2|rr) &= \frac{P(B_2)P(rr|B_2)}{P(B_2)P(rr|B_2) + P(B_1)P(rr|B_1)} \\ &= \frac{0.5 * 0.49}{0.5 * 0.49 + 0.5 * 0.25} \\ &\approx 0.66 \end{aligned}$$

Conservatism = moving opinion in same direction as Bayesian calculation, but less extreme.

Note: This, arguably, is the *opposite* of base-rate neglect.

Other results: *inertia/primacy effect* (initial draws move opinions more); *response bias* (seemed reluctant to deviate from initial experimenter-set probability; no conservatism close to this number); better fit between subjective estimates of sampling distributions and update that objective sampling distribution.

Conclusion from research: Bayesianism an approximate match to people's updating procedures.

Hahn and Harris are fans of this approach.

Worries: artificial. And Kahneman and Tversky came on the scene...

Heuristics and Biases: Find qualitative violations of probability and expected utility theory; take as evidence for useful but error-prone heuristics.

E.g. Conjunction fallacy and representativeness.

Worries: how costly are these errors. And how common? Systematic flaws vs. brain teasers. Grab-bag of heuristics, vs. explanation of how/why we have them?

E.g. some rules that lead to conjunction fallacy can help cancel out noise and so be fairly accurate (Juslin et al. 2009).

Social psychology: show that people's judgments/decisions are sensitive to factors they shouldn't be, e.g. reasoning to make yourself happy or avoid cognitive dissonance.

Mired by contradictory results:

- Better-than-average effect (drivers) vs. worse-than-average effect (calculus; juggling).
- False-consensus effect vs. false-uniqueness effect.
- Self-enhancement bias vs. self-deprecating bias.
- Selective exposure to congenial evidence, vs. selective exposure to incongruent evidence.

Etc...

Introduce "moderators"? Hahn and Harris skeptical: without theoretical motivation, these are non-explanatory epicycles.

II. A Methodology for "bias"

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...

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

First, *expected* deviation. Think of e_X as an estimate of some quan-

tity X.

E.g. e = proportion of people in your sample with covid, and X = proportion of population with covid.
Or e = credence in p and X = truth-value of p.

If X and e_X and are quantities (which you can estimate), so is the *divergence* between the two, $e_X - X$.

The estimate is *biased* when you should expect it to diverge from the quantity: $\mathbb{E}(e_X - X) \neq 0$.

$\mathbb{E}(e_X - X) = \sum_{t \in \mathbb{R}} P(e_X - X = t) \cdot t$

Actually, I now realize, their view is stronger than this. It's a bit weird. I think the Bayesian-friendly way to state it is: for all values t of the quantity X, $\mathbb{E}(e_X | X = t) = t$; or equivalently, $\mathbb{E}(e_X - X | X = t) = 0$.

This gives some implausible verdicts, actually, which we can get into...
Perhaps a **paper topic** here; I don't know the Bayesian stats literature as well as I wish.

(Strangely, Hahn and Harris don't commit to *what* we should calculate the expectation relative to. At times they are inclined towards objective chances or some such.)

An unbiased estimate: your future credence the coin landed heads, after asking whether it did:

Let X be the truth-value of H. If $X = 1$, then $P^+(H) = 1$, and if $X = 0$, then $P^+(H) = 0$. Currently 50-50. So

$$\mathbb{E}(P^+(H) - X) = P(X = 1)(1 - 1) + P(X = 0)(0 - 0) = 0$$

Another: a coin has a 50-50 chance of either being biased 80% in favor of heads or 80% in favor of tails. $X = 1$ if the former (truth value of *heads-based*), 0 if the latter.

$e_X = P^+(\textit{heads-biased})$ = your future credence the coin is biased toward heads, after seeing how it landed on one toss.

So $P^+(hb) = 0.8$ if H, and $P^+(hb) = 0.2$ if T.

$\mathbb{E}(P^+(hb) | X = 1) = 0.8 * 0.8 + 0.2 * 0.2 = 0.68$, and

$\mathbb{E}(P^+(hb) | X = 0) = 0.8 * 0.2 + 0.2 * 0.8 = 0.32$, so:

$$\begin{aligned}\mathbb{E}(P^+(H) - X) &= P(X = 1) \cdot \mathbb{E}(P^+(hb) - X | X = 1) + P(X = 0) \cdot \mathbb{E}(P^+(hb) - X | X = 0) \\ &= P(X = 1)(0.68 - 1) + P(X = 0)(0.32 - 0) = 0\end{aligned}$$

A biased example: fair coin will be flipped. Bill is delusional, so that no matter what he sees, he'll increase his confidence that it landed heads, e_X, to 0.8. $X = 1$ if the coin lands heads, 0 otherwise.

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

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:

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

But if you have to decide whether to bet on H, would you rather decide yourself or let Jill decide on your behalf?

Jill! She'll either have the same opinion as you (0.5 if T), or a more accurate one (1 if H).

Upshot Sometimes biased estimators can (predictably) more accurate than unbiased ones. So, say H&H, bias itself doesn't indicate irrationality.

The example they use to illustrate this relies on their alternative definition of bias.

Consider estimating the proportion of a population that has a given trait based on a sample. Bayesian: update by conditioning, form expectation. Frequentist: just use sample proportion. The former is biased because more conservative; but is expectedly more accurate.

Why? Frequentist jumps to conclusions, which can be bad with small samples. If $5/5$ people sampled have covid, should you estimate the population infection rate at 100%? Frequentist does; Bayesian doesn't.
If b is Bayesian's prior estimate and e_X is their posterior, then if $X = t > b$, $\mathbb{E}(e_X|X = t) < t$.

Upshot: on rigorous definitions of "bias", it's not always detrimental to accuracy (or, in turn, decision).

So to conclude irrationality from bias we need to show that this bias is *detrimental* too:

They bring in "signal detection theory" and how asymmetries in payoffs can lead to "biases" in decisions as well. Makes sense to be biased towards not convicting in criminal trials, because of asymmetry in costs of errors.

Thus to show an irrational bias, must show

1) The estimator must be biased *in expectation*
2) These biased estimates must hold for a wide range of values, so that it's *systematically wrong*.
3) These systematic errors must be shown to be [expectedly?] *costly*.

Relevant for their definition: wide range of values of $X = t$. Not relevant for my variant.

Their bold claim: Few if any of the research on bias has done all of (1)–(3).

They say heuristics and biases fails (3), and maybe (2). Most other research fails (3).

Let's look at some examples!

References

Briggs, R., 2009. 'Distorted Reflection'. *Philosophical Review*, 118(1):59–85.

Hahn, Ulrike and Harris, Adam J.L., 2014. 'What Does It Mean to be Biased. Motivated Reasoning and Rationality.' In *Psychology of Learning and Motivation - Advances in Research and Theory*, volume 61, 41–102.

Juslin, Peter, Nilsson, Håkan, and Winman, Anders, 2009. 'Probability Theory, Not the Very Guide of Life'. *Psychological Review*, 116(4):856–874.

Klayman, Joshua and Ha, Young-won, 1987. 'Confirmation, disconfirmation, and information in hypothesis testing.' *Psychological Review*, 94(2):211–228.

Wason, Peter C, 1960. 'On the failure to eliminate hypotheses in a conceptual task'. *Quarterly journal of experimental psychology*, 12(3):129–140.

Weisberg, Jonathan, 2007. 'Conditionalization, reflection, and self-knowledge'. *Philosophical Studies*, 135(2):179–197.