Predictable polarization

Kevin Dorst 24.223 Rationality

I. Confirmation bias as reflection failures

We saw from Kelly:

- The choice to scrutinize selectively can certainly be Bayesian.
- So long as *failing* to find a flaw *lowers* your credence, the update can be perfectly rational.

Why think this is a form of bias? Because often we can *predict* (or *have expectations for*) how it will shift our beliefs.

What confirmation bias is not:

- Not about being *likely* to raise your credence.
- Not about someone *who knows more than you* being able to predict how your beliefs will shift.

Lottery with 10 tickets; credence you lost?

Me knowing you'll raise your credence that my Dad's birthday is in June.

Proposal: Your inquiry exhibits **confirmation bias** toward q iff your expectation of your updated credence in q is higher than your prior:

$\mathbb{E}_P(P^+(q)) > P(q)$. I.e. violate the expectation-version of Reflection.

It's definitely possible to search for evidence for q without exhibiting confirmation bias toward q.

Example: word searches.
Possibilities = (n, c, f): no word, $c$ompletable but don't find, and $f$ind.

$$P = \begin{pmatrix} n & c & f \\ \hline \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \end{pmatrix}, \text{ while } P^+ = \begin{pmatrix} 2/3 & 1/3 & 0 \\ 2/3 & 1/3 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

P is stationary wrt P^+: $\mathbb{E}_P(P^+) = P$.

Salow: selective scrutiny is rational only if it satisfies Reflection.

$\mathbb{E}_P(P^+(q)) = P(q)$.

Dorst: the problem of polarization is that we violate Reflection.
Examples: Googling symptoms; biased or one-sided sources; Pascal's Wager; lawyer's argument; going to college.

II. Is confirmation bias irrational?

An update (P, P^+) satisfies the **value of evidence** iff, for all decisions[1], P expects the option that P^+ recommends[2] to be better than the option P recommends.
$\rightarrow$ P wants to give power of attorney to P^+.

[1] Set of options $\{X_1, ..., X_n\}$, fxns from worlds w to utilities $X_i(w) \in \mathbb{R}$
[2] The X_i that maximizes $\mathbb{E}_{P^+}(X_i)$

$\Rightarrow$ P expects P^+ to be more accurate than P

Fact: there are updates that satisfy the value of evidence that exhibit confirmation bias. *Example:*

$$P = (\tfrac{1}{2}, \tfrac{1}{4}, \tfrac{1}{4}), \text{ while } P^+ = \begin{pmatrix} 2/3 & 1/3 & 0 \\ 1/3 & 2/3 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

$\mathbb{E}_P(P^+) = (\frac{5}{12}, \frac{4}{12}, \frac{3}{12})$, so
$\mathbb{E}_P(P^+(word)) = \frac{7}{12} \approx 0.58 > 0.5 = P(word)$

How? This is possible because P^+ has *higher-order uncertainty:* it is unsure of its own values.
$\rightarrow$ an *ambiguity-asymmetry*.

The rational response to *not-finding* is to be unsure what the rational response is.

Claim 1: Since this update satisfies the value of evidence, it can be rational despite exhibiting confirmation bias.

Claim 2: Repeating this (rational) process can lead to predictable, profound polarization.

References

Henderson, Leah and Gebharter, Alexander, 2021. 'The role of source reliability in belief polarisation'. *Synthese*, 1–23.
Jern, Alan, Chang, Kai Min K., and Kemp, Charles, 2014. 'Belief polarization is not always irrational'. *Psychological Review*, 121(2):206–224.
Nickerson, Raymond S., 1998. 'Confirmation bias: A ubiquitous phenomenon in many guises.' *Review of General Psychology*, 2(2):175–220.
Salow, Bernhard, 2018. 'The Externalist's Guide to Fishing for Compliments'. *Mind*, 127(507):691–728.