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.

Salow: selective scrutiny is rational only if it satisfies Reflection.
$\rightarrow$ *Cases:* creationist arguments; defense attorneys.

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

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

Example: searching for evidence.
Possibilities = (n, e, f): (1) no evidence; (2) $e$xists, don't find; (3) $f$ind.

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

P is stationary wrt P^+: $\mathbb{E}_P(P^+) = P$. Salow says all searches for evidence should look like this, on pain of irrationality.

Past-Kevin: real people violate Reflection all theme.

Examples: Googling symptoms; biased or one-sided sources; Pascal's Wager; lawyer's argument; going to college; Cohen and graduate school.

II. Against irrationalism

Suppose this is right. Should we conclude that our (predictable) polarization is irrational?

Past-Kevin gives a quick argument that we shouldn't:

P1 (No Special Pleading) If you think polarization is irrational, you should think that *your* political beliefs were formed irrationally.

P2 (Anti-Akrasia) If you should think that your political beliefs are irrational, then you should give them up.[1]

P3 (Resoluteness) You *shouldn't* give up your political beliefs.

C You shouldn't think polarization is irrational.

How should irrationalists resist this argument?

[1] Suspend judgment or have middling credence

III. *Rational* confirmation bias? A warm-up

Professor Polder is a polarizing figure. Has fans and critics.

Standard Bayesianism: known prior P, fixed question (partition) Q, posterior P^+ obtained by conditioning P on the true answer to Q.
$\rightarrow$ You're always answering the same (set of) question(s).

That's nuts. We have *limited attention*.

Basic idea: which question you ask (what you notice) is correlated with what you see; this leads people to be *selectively sensitive* to information.

Return to Polder. Suppose can ask two questions about each comment:

- Q_a = *Is it aggressive?* Possible answers: $\{A, \neg A\}$
- Q_i = *is it insightful?* Possible answers: $\{I, \neg I\}$

Crossing these, there are 4 possible options. $(A\neg I \quad AI \quad \neg AI \quad \neg A\neg I)$

Suppose in each comment he's 50%-likely to be aggressive, 50%-likely to be insightful, and the two are independent. Calibrated priors:

$$F = C = \left(\begin{array}{cccc} A\neg I & AI & \neg AI & \neg A\neg I \\ \hline 0.25 & 0.25 & 0.25 & 0.25 \end{array}\right)$$

Fans and critics have different habits of mind—which question they ask (so what they notice) depends on what they see. Suppose:

- **Fans:** When his comments are insightful, fans always ask Q_i, noticing that they're insightful. When his comments are not-insightful, they always ask Q_a—noticing *whether* they're aggressive or not.
- **Critics:** When his comments are aggressive, critics always ask Q_a, noticing that they're aggressive. When his comments are not-aggressive, they always ask Q_i—noticing *whether* they're insightful or not.

We can formalize their posteriors again using a matrix. Eg the fans':

$$F^+ = \left(\begin{array}{c|cccc} & A\neg I & AI & \neg AI & \neg A\neg I \\ \hline A\neg I & 0.5 & 0.5 & 0 & 0 \\ AI & 0 & 0.5 & 0.5 & 0 \\ \neg AI & 0 & 0.5 & 0.5 & 0 \\ \neg A\neg I & 0 & 0 & 0.5 & 0.5 \end{array}\right)$$

So:
If $A\neg I$, condition on A, so $F^+(I) = 0.5$.
If AI or $\neg AI$, condition on I, so $F^+(I) = 1$.
If $\neg A\neg I$, condition on $\neg A$, so $F^+(I) = 0.5$.

Notice: Despite always conditioning on truths, fans' opinions about whether he's insightful are biased: if I, jump to 100%-confident; if $\neg I$, stay at 50%-confident $\rightsquigarrow$ on average, increase confidence.

Average $F^+(I)$ is $0.5 \cdot 1 + 0.5 \cdot 0.5 = 75\%$

Similarly for critics' posteriors:

$$C^+ = \left(\begin{array}{c|cccc} & A\neg I & AI & \neg AI & \neg A\neg I \\ \hline A\neg I & 0.5 & 0.5 & 0 & 0 \\ AI & 0.5 & 0.5 & 0 & 0 \\ \neg AI & 0 & 0.5 & 0.5 & 0 \\ \neg A\neg I & 0.5 & 0 & 0 & 0.5 \end{array}\right)$$

So:

If $A\neg I$ or AI, condition on A, so $C^+(A) = 1$.

If $\neg AI$, condition on I, so $C^+(A) = 0.5$.

If $\neg A\neg I$, condition on $\neg I$, so $C^+(A) = 0.5$.

Biased toward A: if A, jump to 100%-confident; if $\neg A$, stay at 50%.

Average $C^+(A)$ is $0.5 \cdot 1 + 0.5 \cdot 0.5 = 75\%$

Because of these biases, iterating this process **leads them to polarize**.

Is this *rational?*

Rationality as Value: An update is rational if it increases your accuracy and improves your decisions *about the questions you care about*.

Notice: if Fans only care about Q_i, the update is valuable.

Their credence in the answers, $\{I, \neg I\}$, will either stay the same (if $\neg I$) or become perfectly accurate (if I).

Similarly for Critics about Q_a.

Problem: it doesn't seem that Fans/Critics can be *self-aware*.

If they were, they'd *realize* that they're being selective, and correct for it.

But what if we could construct a case where they *can't* correct for it?

IV. Ambiguity Asymmetries

Sometimes our evidence is *clear*. Other times it's *ambiguous*.

Let P be your (rational) credences. Then:

- Your opinion about q is *clear* if you have *higher-order certainty*: you're sure what your (rational) credences are.
- Your opinion about q is *ambiguous* if you have higher-order *un*certainty: you're unsure what your (rational) credences are.

For some x, $P(P(q) = x) = 1$.
Eg: a fair coin lands heads?

For all x, $P(P(q) = x) < 1$
Eg: more than a dozen spoons?

What if we could construct a case where evidence is *asymmetrically ambiguous*?
$\rightarrow$ You can get clear evidence one way, but only ever ambiguous evidence the other way.

Example(?): word searches. Theory:

Let's try it.

You're either *g*ood or *b*ad at word-searches:

- If g, you'll find a word if there is one.
- If b, you won't find a word even if there is one.
- Either way, there's a 50-50-chance to be a word.

$P(w|g) = P(w|b) = 0.5$

Your opinion about whether you're good or bad is ambiguous: you're unsure whether you're 2/3- or 1/3-confident that you're good. Prior:

Focusing just on g and b: $P_{mod\ \{g,b\}} = \begin{pmatrix} g & b \\ \hline 2/3 & 1/3 \\ 1/3 & 2/3 \end{pmatrix}$

Crossing this with word (w) vs. not ($\overline{w}$): $P = \begin{pmatrix} gw & g\overline{w} & bw & b\overline{w} \\ \hline 2/6 & 2/6 & 1/6 & 1/6 \\ 2/6 & 2/6 & 1/6 & 1/6 \\ 1/6 & 1/6 & 2/6 & 2/6 \\ 1/6 & 1/6 & 2/6 & 2/6 \end{pmatrix}$

So $P(w) = 0.5$ and $P(P(w) = 0.5) = 1$.

Suppose you're at gw, so your actual prior is $P_@ = (2/6, 2/6, 1/6, 1/6)$

Update by conditioning on whether you found a word or not.
Notice: find = $\{gw\}$, while $\neg find = \{g\overline{w}, bw, b\overline{w}\}$.

So your posterior, in each world, is [»]:

$P^+ = \begin{pmatrix} gw & g\overline{w} & bw & b\overline{w} \\ \hline 1 & 0 & 0 & 0 \\ 0 & 1/2 & 1/4 & 1/4 \\ 0 & 1/5 & 2/5 & 2/5 \\ 0 & 1/5 & 2/5 & 2/5 \end{pmatrix}$

If you find a word, your evidence is clear.
If you don't find a word, it's ambiguous.

- If g, $\neg$*find* is strong evidence against there being a word.
- If b, $\neg$*find* is weak evidence against there being a word.

$P^+(w|\neg f) = 1/4 = 0.25$

$P^+(w|\neg f) = 2/4 = 0.4$

But $\neg$*find* also *provides evidence that you're bad at searching*, i.e. that you wouldn't find a word even if there was one.

I.e. $\neg$*find* provides evidence that it's only *weak* evidence against *word*.

- If g, then $P(b) = 1/3 = 0.33 < 0.5 = 1/2 = P(b|\neg f)$.
- If b, then $P(b) = 2/3 = 0.67 < 0.8 = 4/5 = P(b|\neg f)$.

So the update leads to **expectable polarization**:
$\mathbb{E}_{P_@}(P^+(w)) = 0.55 > 0.5 = P_@(w)$.

$= \frac{2}{6}(1) + \frac{2}{6}(0.25) + \frac{1}{6}(0.4) + \frac{1}{6}(0.4)$
whereas
$\frac{2}{6}(1) + \frac{2}{6}(0.25) + \frac{1}{6}(0.25) + \frac{1}{6}(0.25) = 0.5$

And—unlike our selective-attention model—this happens through dynamics *you're aware of*. The problem is that you don't know exactly what your credences are, so you can't correct for it.

V. Predictable Polarization

Iterating, our Bayesian agents will predictably go off the rails.

Repeat this with a bunch of IID word-searches, $w_1, ..., w_n$.

Since, the expectation of your posterior is 0.55, then[2], you're initially quite confident that your average posterior will be roughly 0.55.

[2] By the law of large numbers

Mean posterior equals posterior mean: if you're average posterior is ≈ 0.55, your posterior estimate for the proportion of true w_i is ≈ 0.55.

And since they're independent, that implies that you'll be *very confident* that roughly 55% of the w_i are true, and so confident that *more than 50%* are.

Polarization? Headsers and Tailsers. Our searches pull in opposite directions, so we'll polarize over whether more than 50% of the coins landed heads.