

Confirmation Bias (Chapter 8)

Kevin Dorst

24.805, Fall 2025

I. What Confirmation Bias Is Not

Standard definition (Nickerson 1998): ‘Seeking or interpreting evidence in ways that are **partial** to existing beliefs.’

What does ‘partial’ mean?

Caricature: People update on congruent & reject incongruent evidence.

Leading, eg, to the *backfire effect*.

Claim: This isn’t supported. Rather:

Arguments Work: When presented with arguments for a claim, people (on average) shift their beliefs toward it, regardless of their priors.

What is it to construct an ‘**argument**’ for q , here? It’s a search for evidence favoring q .

An attempt to find e such that $P(q|found\ e) > P(q)$.

Sometimes people *should* ‘backfire’ in response to an argument—people should update *relative to their expectations*.¹

¹ Lawyer: ‘My client is so friendly.’

→ They do (e.g. Melnikoff and Strohminger 2024)

If they *systematically* backfired against arguments that don’t fit with their views, that’d be suspicious.

→ But they don’t. *Average* backfires are rare.

Strong evidence for ‘persuasion in parallel’.

Coppock 2023

Further empirical findings:

We Persuade Ourselves: When people construct *their own* arguments for q , they tend (on average) to increase their credence in q .

I.e. ‘self-persuasion’; Schwardmann et al. 2022

Motives Work Indirectly: People’s motives do influence their beliefs—but only indirectly, through search for evidence favoring those beliefs.

I.e. ‘motivated reasoning’; Kunda 1990; Williams 2023

II. What Confirmation Bias Is

Why not say that ‘confirmation bias’ just Standard-Bayesian updating + differing background beliefs?

Jern et al. 2014; Gershman 2019; Benoit and Dubra 2019; Henderson and Gebharter 2021

Unsatisfying. Confirmation bias was meant to explain *failures* of convergence. Famously, Standard Bayesians converge to the truth.

Puzzle: if people are responsive to evidence, why don’t they converge?

At least on factual matters with plentiful evidence: *Who does more chores? or Is climate change made worse by humans?*
Lottery example with 10 tickets

Well, why *wouldn’t* they converge if updated like the caricature?

What about *softened version* of the caricature: 50% likely to update if learn $\neg win$, 100%-likely if learn win .

→ In both cases, asymmetric sensitivity means that new evidence would

(on average) *skew* your assessment of the truth, violating:

Martingale ('Reflection'): No matter how you search for evidence, your expectation for your posterior should match your prior.

Martingale is valid on all Standard-Bayes models.

→ Eg ask a Kevin-partisan about how many lotteries I won.

Martingale is essential for convergence. When it fails, *even if things go exactly as your prior expects*, your posterior will be thrown off the rails.

Notice something odd: arguments *predictably* work.

→ Seems to violate Martingale.

Proposal: You exhibit *confirmation bias* on q iff you search for evidence in a way that *predictably, on average* increases your probability for q .

Empirically: People seem to violate Martingale.

- Seems like people know implicitly that Arguments Work, eg how Motives Work Indirectly.
- **Experiment 1**: people seem to violate Martingale in self-persuasion design. Randomized to Pro/Con conditions.
 - t_1 : Elicit *prior* and *estimated posterior*;
 - t_2 : Write arguments for [/against] claim.
 - t_3 : Elicit *posterior*.

(Main-1) People should persuade themselves.

(Main-2) People should expect to persuade themselves (MFs)

(Main-3) People's expectations should be reasonably calibrated.

III. Why Seeking Confirmation Works

Why *doesn't* it work for Standard Bayesians? Because (1) they update relative to their expectations, and (2) *their expectations are clear*.

Idea: searching for evidence under ambiguity leads to *ambiguity asymmetries*—avg differences in how ambiguous your posteriors are when the search is completable than when it's not.

Eg **word searches**:

- If you *find* (f) a word, your posterior that it's *completable* (c) is clear.
- If you *don't find* ($\neg f$), your posterior in c , $P^+(c)$, is ambiguous. You update by conditioning, so $P^+(c) = P(c|\neg f)$. But what is *that*? By Bayes theorem:

$$P^+(c) = \frac{1 - P(f|c)}{2 - P(f|c)}$$

But what did you expect?—What was $P(f|c)$? It was ambiguous. So your posterior $P^+(c)$ is ambiguous too.

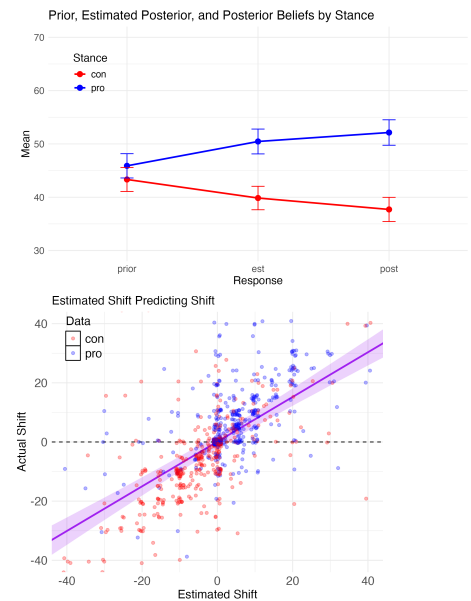
$\mathbb{E}_{P_a}(P^+(q)) = P_a(q)$
 $\mathbb{E}_{P_a}(P^+(q))$ is an avg of the possible values of $P^+(q)$, weighted by prior P_a .

Always says strongest $X \geq x$ he can

If I tell you I'll argue for q , then it seems that *before hearing the argument*, you can expect it to increase your probability for q .

I.e. iff you violate Martingale

$$\mathbb{E}_{P_a}(P^+(q)) > P_a(q)$$



Priors were certain of how much they expected the evidence:
 if $P_a(e) = x$, then $P_a(P(e) = x) = 1$

You know that $P^+(c) = P(c|f) = 1$

$P(f|c)$ = how good you expected to be at word searches

There's no x such that $P(P(f|c) = x) = 1$
 No y such that $P^+(P^+(c) = y) = 1$

Upshot: There's an *ambiguity asymmetry* in search. If you find a word, you know what your posterior in c is; if you don't, you're unsure.

This can drive Martingale failures.

Toy model:

- Either you think you're good ($P = \pi_g$) or bad ($P = \pi_b$).
- Either way, c is 50%-likely.
- What you think is ambiguous:
 If you think you're good, you're $2/3$ -confident of that.
 If you think you're bad, you're $2/3$ -confident of that.
- Suppose you'll *find* iff *both* completable *and* you think you're good.
- What happens?

Suppose, in fact, $P = \pi_g$.

$$\pi_i(c|P = \pi_g) = 0.5 = \pi_i(c|P = \pi_b)$$

$$\pi_g(P = \pi_g) = 2/3, \text{ so } \pi_g(P = \pi_b) = 1/3$$

$$\pi_b(P = \pi_b) = 2/3, \text{ so } \pi_b(P = \pi_g) = 1/3$$

$$\text{So } f = \{g_c\}, \text{ and } \neg f = \{g_{\bar{c}}, b_c, b_{\bar{c}}\}$$

Notation:

• A credence function is a vector, so $(.33, .33, .17, .17)$ is 0.66 in $\{g_c, g_{\bar{c}}\}$

• Row i = credence function at world i

• P is prior, P^+ is posterior

$$P \approx \begin{pmatrix} & g_c & g_{\bar{c}} & b_c & b_{\bar{c}} \\ g_c & .33 & .33 & .17 & .17 \\ g_{\bar{c}} & .33 & .33 & .17 & .17 \\ b_c & .17 & .17 & .33 & .33 \\ b_{\bar{c}} & .17 & .17 & .33 & .33 \end{pmatrix} \quad P^+ = \begin{pmatrix} & g_c & g_{\bar{c}} & b_c & b_{\bar{c}} \\ g_c & 1 & 0 & 0 & 0 \\ g_{\bar{c}} & 0 & .50 & .25 & .25 \\ b_c & 0 & .20 & .40 & .40 \\ b_{\bar{c}} & 0 & .20 & .40 & .40 \end{pmatrix}$$

- If you have π_g , not-finding is *strong* evidence.
- If you have π_b , not-finding is *weak* evidence.
- But you're more likely to not-find if (you think) you're bad at this!
 → So not-finding is also evidence that you had π_b , i.e. that *you didn't expect to find a word*.

$$\pi_g(c|\neg f) = 0.25.$$

$$\pi_b(c|\neg f) = 0.40$$

Martingale failure: $\mathbb{E}_P(P^+(c)) \approx 0.55$.

→ Due to *asymmetric* accuracy improvements.

Dampening disconfirming effect of $\neg f$.
 $\approx .33(1) + .33(.25) + .33(.40) = .55$

If c , avg $P^+(c) = 0.8(+0.3)$

If $\neg c$, avg $P^+(c) = 0.3(-0.2)$

Overall average: 0.55

No accident. **Liu's Theorem:** Given nontrivial ambiguity, there's always partition s.t. conditioning would violate Martingale.

In natural class models, MF tends to be in direction of search.

More sources of ambiguity if you fail to *find*, since finding requires many things to go right.

Driven by 'ambiguity asymmetry': less ambiguity if *find* than if \neg *find*, so on avg less ambiguity if c than if $\neg c$.

Idea: Arguing for q is like searching for a word.

Word-completion tasks

The proportion of completable strings.

Your estimate of the proportion.

Cognitive search for a word.

Finding a word (clear).

Failing to find a word (amb).

Increase accuracy about each task, but *asymmetrically*.

Shift estimates in the direction of search.

Arguments

The balance of evidence.

Your strength of belief in q .

Construction of argument for q .

Finding strong evidence for q (clear).

Failing to find strong evidence (amb).

Increase accuracy about each bit of evidence, but *asymmetrically*.

Shift beliefs in the direction of arguments.

IV. Predictions

1) Martingale should fail in direction of search.

2) [Pilot] Clarity will be higher when shifting up than down

Similar design; only Pro.

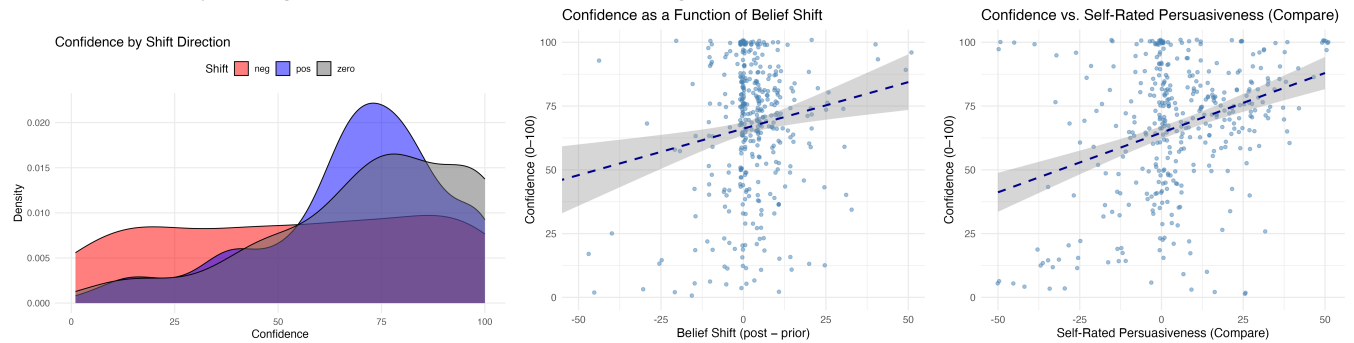
t_1 : Elicit *prior*.

t_2 : Write arguments for claim.

t_3 : Elicit *posterior, comparison* to expectations,² and *clarity* of response.³

(Main-2) Clarity is higher with positive than negative shift.

(Main-3) Clarity is higher with better than worse arguments.



It does—Experiment 1

Since amb-asymmetries drive shifts

Pilot on Prolific. 98 subjects after comp/attn checks. 4 iterations each.

² 'How good were the arguments you generated compared to what you expected beforehand?'

³ 'How confident you are that you responded to the evidence (the arguments you generated) properly?'

3) Asymmetry in ambiguity, rather than strength drives shifts

Either get strong or weak evidence about coin.

Pro vs. Con: possible to get strong evidence if *heads* vs. if *tails*

Amb. vs. Clear: do they do it with word-searches or draws from a bag of known composition?

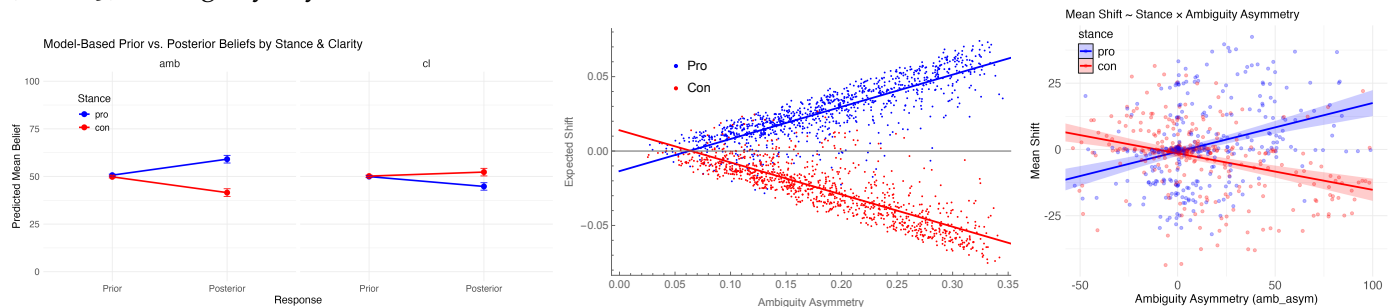
Measure confidence used evidence properly; ambiguity-asymmetry as difference in confidence when evidence was 'available'⁴ vs. not.

⁴ Completable if amb; bag contains definitive marble if clear

(Main-1) Clarity predicts ambiguity-asymmetry

(Main-2) Will shift in direction of search if amb; not if clear.

(Main-3) Ambiguity-asymmetries will correlate with shifts.



References

- Benoît, Jean Pierre and Dubra, Juan, 2019. 'Apparent Bias: What Does Attitude Polarization Show?' *International Economic Review*, 60(4):1675–1703.
- Coppock, Alexander, 2023. 'Persuasion in parallel'. In *Persuasion in Parallel*. University of Chicago Press.
- Gershman, Samuel J., 2019. 'How to never be wrong'. *Psychonomic Bulletin and Review*, 26(1):13–28.
- Henderson, Leah and Gebharter, Alexander, 2021. 'The role of source reliability in belief polarisation'. *Synthese*, 199(3-4):10253–10276.
- Jern, Alan, Chang, Kai Min K., and Kemp, Charles, 2014. 'Belief polarization is not always irrational'. *Psychological Review*, 121(2):206–224.
- Kunda, Ziva, 1990. 'The case for motivated reasoning'. *Psychological Bulletin*, 108(3):480–498.
- Melnikoff, David E. and Strohminger, Nina, 2024. 'Bayesianism and wishful thinking are compatible'. *Nature Human Behaviour*, 8(4):692–701.
- Nickerson, Raymond S., 1998. 'Confirmation bias: A ubiquitous phenomenon in many guises'. *Review of General Psychology*, 2(2):175–220.
- Schwardmann, Peter, Tripodi, Egon, and van der Weele, Joël J., 2022. 'Self-Persuasion: Evidence from Field Experiments at International Debating Competitions'. *American Economic Review*, 112(4):1118–1146.
- Williams, Daniel, 2023. 'The case for partisan motivated reasoning'. *Synthese*, 202(89).