

Reasonable Convergence: Mandelbaum 2019 and Chapter 1

Kevin Dorst (kmdorst@mit.edu)

24.805, Fall 2025

I. Mandelbaum, Troubles with Bayesianism

Descriptive Bayesianism.

- Dominant in low-level cognition (vision, motor control); increasingly popular for high-level cognition (causal inference; belief updating).
- Marr's levels: computational, algorithmic, implementational.
→ Bayesians focus on computational & algorithmic
- Need to distinguish *competence* from *performance* (Cohen 1981).

Shape of bird's wing.
Vision: recover 3D map from 2D array.
Neuroscientists and microprocessors
So Mandelbaum thinks heuristics-and-biases program isn't decisive.

Three problems for (descriptive, "Imperial" Bayesianism):

Belief perseverance. Firefighters and risk attitudes.

Even those subjects who merely contemplated a hypothetical relationship between risk-seekingness and firefighting would stubbornly adhere to that belief when confronted with (fictional) mounds of data that seemed to conclusively show a contradictory link between risk-aversion and firefighting. (147)

- M doesn't think this is definitive. Why? Presumably because these studies only show a *statistically-detectable difference* in the groups' opinions, after debriefing.

Groups always update toward the info contained in the debrief. In general, as we'll see (Ch. 8), *arguments work*

Biased assimilation. Partisans and capital-punishment studies.

- Conflicting evidence \rightsquigarrow selective scrutiny \rightsquigarrow more disagreement.
- Again, M doesn't think definitive. Presumably because (1) increasing disagreement is due to differential updating *toward the evidence*, and (2) Bayesian models can explain this.

Eg Jern et al. 2014; Henderson and Gebharter 2021.
Increase proselytizing

Belief disconfirmation. August Petermann (Arctic); millennial cults.

- Batson 1975: Religious teenagers read article; told was 'denied publication at request of the World Council of Churches, because of the obvious crushing effect it would have on the entire Christian world'.

Of the subjects who believed that $s = \textit{Jesus was the son of God}$:

- If they thought article untrue¹, belief in s went down a bit or stayed the same.
- If they thought article true², their belief in s went *up*.

¹ 3,4, or 5 on a 1–5 agreement scale to 'The article is untrue'

² 1 or 2 on a 1–5 agreement scale for 'The article is untrue'

These participants both believed that s (that Jesus was the son of God) and also agreed that they just received evidence that not- s . Like Petermann and the millennial cultists before them, these subjects increased their belief that s in the face of evidence that they took to be disconfirming. (150–1)

Other cites: Plous 1991; Liberman and Chaiken 1992; McHoskey 1995; Munro and Ditto 1997; Taber and Lodge 2006

Theory: the *psychological immune system* (PIS)

- Beliefs are governed by a psychological immune system who's *proper function* is to protect against threats to your self-identity.
→ getting evidence against a core belief induces dissonance.
→ people change their attitudes to escape psychological discomfort.

- Only those who accepted article as true were put in dissonant state, so only their PIS's were activated.
- Can also explain biased-assimilation and belief perseverance.

So (only) they increased belief in s .

Thoughts? Questions? Evaluations?

II. The Assumption of Convergence

Reasonable Convergence (RC): When evidence is plentiful, if people take reasonable steps to figure out the truth, they'll usually succeed.

- 'Reasonable' = epistemic (not practical) rationality.
- RC + polarization \Rightarrow many people are being unreasonable

Ch. 12 will argue that this inference exacerbates political animosities

Main Thesis: Due to ambiguity, RC is false. (Even for ideal Bayesians.)

What are *Bayesians*?

- Mathematical models of beliefs.
- Assign fine-grained degrees of confidence ('**subjective probabilities**' or '**credences**'), from 0–100%, to propositions ('claims', 'events').
- Treat things as evidence for things that make them likely. That is: treat e as evidence for (against) h iff prior thought that h would make e more (less) to be expected.
- 'Bayesian convergence theorems' are often taken to establish RC.

Why is Mandelbaum's argument persuasive? Hypothesis:

- Because people know that Bayesians are supposed to obey Reasonable Convergence.
- But they also know that people *don't* converge, given plentiful evidence—at least not about topics that are important to them, like their self-image (*I do plenty of chores*) or political identity (*I'm not racist/sexist/etc.*) or core beliefs (*God does(n't) exist*).
- So evidence like Mandelbaum's³ offers an explanation for *why* people don't converge.

"A Bayesian mind is, at its core, a rational mind."

³ or Meehl's, or Kahneman & Tversky's

Alternative explanation: **Clarity and Ambiguity**

Examples: coin vs. word-search; pi vs spoons

\rightarrow Our clear judgments are reliable and confident.

\rightarrow Our ambiguous ones are noisy and doubtful.

Hypothesis: under ambiguity, we *aren't* sure what our subjective probabilities are, but we *still have* them:

Think of $P(q)$ like your blood pressure.

Ambiguity as Higher-Order Uncertainty:

- Your judgment about q is **clear** when you're certain of what subjective probability is—i.e. you have higher-order certainty.
- Your judgment about q is **ambiguous** when you're *uncertain* of what your subjective probability is—i.e. you have higher-order *uncertainty*.

' P_a ' = a constant for your actual prior.
' P ' = a random variable that picks out your prior P_w in various worlds w .

For some x , $P_a(P(q) = x) = 1$.

For all x , $P_a(P(q) = x) < 1$.

Genuine higher-order uncertainty vs. (mere) probabilistic uncertainty.

Mathematical facts:

M.1 Almost all Bayesian models—including those used to prove RC—implicitly presuppose clarity.

M.2 The same models, under ambiguity, lead to bias and often polarization in the face of plentiful evidence.

Hurdles to believing that the mathematical facts matters:

- 1) Conceptual/normative/mathematical qualms with HOU.
→ Chapters 2–6 will address these.
- 2) Empirical work on human foibles—surely people can't even approximate Bayesian reasoning!

Two faces of human cognition

Mechanistic psychology vs. rational psychology.

Anti-Akrasia argument against mechanism ('irrationalism')

Claim: accepting irrationalism is unstable from first-person perspective:

- *Anti-akrasia:* If your estimate of the ideally-rational credence to have in q , given your evidence, should be x , then your credence in q should be x .
- Assume I can reasonably be confident of a predictably-polarized⁴ belief—say, $q = \text{ICE agents shouldn't be allowed to wear masks}$.
- Suppose—for reductio—I should be confident of irrationalism. Then I should think unreasonable steps⁵ led to my confidence in q .
- So I should be have fair confidence—say, at least 0.5—that *if I had ideally responded to my evidence, I would be much less confident in q* .
- But that implies that my estimate for the ideally-rational credence in q is *lower* than 0.95—violating Anti-Akrasia. Contradiction.

Assuming *anti-akrasia*, the options:

- 1) *Reduce confidence in predictably-polarized beliefs.*
→ Do you think that's the right response for your beliefs about politics, religion, and who does more chores?
- 2) *Think I am an exception to mechanistic psychology's explanations.*
→ The empirical evidence is resounding that people of all political stripes, intelligence levels, etc. are influenced by these biases.
- 3) *Be doubtful of irrationalism, even for predictably-polarized beliefs.*
→ But then how can we explain people's failures of convergence?

My theory gives us a way to do (3), doubt irrationalism.

- Rational psychologists are right that we approximate Bayesianism.
- For topics where we can achieve sufficient clarity⁶ we approximate Standard-Bayesians and so converge. That explains our feats.
- For topics where ambiguity is endemic⁷, even approximating Bayesians leads us to polarize.

Ch. 3, in brief. Chs. 4–6, at length.

Why? SB essentially uses your (clear) priors as a fixed standard against which to judge evidence.

→ Without it, hindsight bias is inevitable, sometimes leading to confirmation bias and thereby polarization.

↪ Puzzling double-image

Will do more carefully in Chapter 11

If $\mathbb{E}_{P_a}(\mathcal{P}(q)) = x$, then $P_a(q) = x$, where \mathcal{P} is ideal, P_a is your (actual) reasonable probabilities, and \mathbb{E}_{P_a} are your (mathematical) *expectations*.

⁴ A belief clearly influenced by the biases of mechanistic psychology.

Say $P_a(q) \geq 0.95$.

⁵ Like biased assimilation, belief disconfirmation, and confirmation bias.

$P_a(\mathcal{P}(q) < 0.9) \geq 0.5$.

$\mathbb{E}_{P_a}(\mathcal{P}(q)) < 0.5(0.9) + 0.5(1) = 0.95$

Generally: if I'm y -confident that if I'd been ideally rational I'd be at most x -confident of q , I can be at most

$y(x) + (1 - y)$ -confident of q :

$y = 0.5, x = 0.6 \rightsquigarrow 0.8$

$y = 0.8, x = 0.6 \rightsquigarrow 0.68$

$y = 0.9, x = 0.55 \rightsquigarrow 0.595$

Set aside the detailed psych evidence. People of all stripes don't converge to the truth!

⁶ Well-constrained domains like vision and language; some parts of science

⁷ Politics, religion, departmental dramas, (sometimes) household chores

References

- Batson, C. Daniel, 1975. 'Rational processing or rationalization? The effect of disconfirming information on a stated religious belief.' *Journal of Personality and Social Psychology*, 32(1):176-184.
- Cohen, L. Jonathan, 1981. 'Can human irrationality be experimentally demonstrated?' *Behavioral and Brain Sciences*, 4(3):317-331.
- 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.
- Liberman, Akiva and Chaiken, Shelly, 1992. 'Defensive processing of personally relevant health messages'. *Personality and Social Psychology Bulletin*, 18(6):669-679.
- McHoskey, John W., 1995. 'Case Closed? On the John F. Kennedy Assassination: Biased Assimilation of Evidence and Attitude Polarization'. *Basic and Applied Social Psychology*, 17(3):395-409.
- Munro, Geoffrey D and Ditto, Peter H, 1997. 'Biased assimilation, attitude polarization, and affect in reactions to stereotype-relevant scientific information'. *Personality and Social Psychology Bulletin*, 23(6):636-653.
- Plous, Scott, 1991. 'Biases in the assimilation of technological breakdowns: Do accidents make us safer?' *Journal of Applied Social Psychology*, 21(13):1058-1082.
- Taber, Charles S and Lodge, Milton, 2006. 'Motivated Skepticism in the Evaluation of Political Beliefs'. *American Journal of Political Science*, 50(3):755-769.