Mandelbaum 2018: The Psychological Immune System

Kevin Dorst

24.223 Rationality

I. The Aim

Descriptive Bayesianism is on the rise. Imperial vs. Local.
→ Objects to Imperial, and some versions of Local.

All vs. some mental processes Bayesian.

Using Marr's levels, focuses on *algorithmic* Bayesianism rather than computational.

Q: Why? Computational is about *goals* of the system, and his alternative has non-Bayesian goals.

Need to show that the deviation from Bayesianism is in the *design* of the system, rather than just being a performance error.

Want to find a feature that *no* Bayesian approximator would have.

Three possibilities:

- Learning blindness: fail to learn what you should.
- Belief persistence: fail to update your beliefs when you should.
- Learning perversity: receiving evidence for $\neg P$ leads you to increase your confidence in P.

II. Learning Blindness and Belief persistence

Belief persistence: the debriefing paradigm.
Get feedback on ability to do some task, then told the feedback was bunk, and maintain (some) belief in direction that feedback pointed.

Notably, though Mandelbaum doesn't mention this, in the classic experiments people *do* become much less confident—it's just that they don't completely revert to their prior.

Mandelbaum's examples: hypotheses about firefighters and risk tolerance; abstract causal learning example.

Unclear to me exactly what Mandelbaum thinks the upshots of these examples are. Since he doesn't press them, presumably he thinks these *could* be exhibited by Bayesian approximators?

They can.

III. Biased Assimilation

Mandelbaum gives a pessimistic reading of the confirmation bias literature.

But seems to acknowledge that Kelly 2008 or Jern et al. 2014-style explanations can make this sort of phenomenon consistent with Bayesian approaches.

Jern: if (1) people believe there is publication bias, and (2) they experience the "false uniqueness effect", then biased updating to mixed evidence is Bayesian.
Kelly: confirmation bias is driven by *selective scrutiny*; and it makes sense to use your priors to figure out what to scrutinize.

Q Why would belief disconfirmation effect pose a *bigger* problem?

IV. Belief Disconfirmation Effect

Here, says Mandelbaum, is the real problem: often people get information that they acknowledge is evidence against P, and yet increase their confidence in P.

Example of Festinger's and other's work following cults. Often predict the end of the world is day d, and then when it doesn't end they ramp up proselytizing afterwards.

Worry 1: These are cults!

Selection effect for irrationality and/or bizarre background beliefs.

Worry 2: In what sense exactly are they "doubling down" on their beliefs? Before they believed the world would end on day d. They clearly *don't* belief *that* anymore—rather they now believe (and proselytize) that the world will end on day $d + n$.

Mandelbaum thinks this generalizes, looking at Batson 1975.

- 50 high school youth group members.
- Divide into those who do (42) and don't (8) believe Jesus is the son of God.
- Present article claiming to show New Testament was fabricated.
- Ask whether they believe the article or not, and then probe their belief that Jesus is the son of God.
- *Results:* Those who rejected article kept their belief the same. Of those who accepted it as true (11 of 42), their average confidence went up.

"denied publication in NYT by request of World Council of Churches, because of devastating impact on Christian community"

A small but statistically significant amount. $4.07 \rightsquigarrow 4.30$ on a 1–5 scale.

This is (supposed to be) a problem because it's updating contrary to what they acknowledge about the evidence:

> There are two morals worth highlighting from the belief disconfirmation effect. The first is that the effect is anathema to any Bayesian model; one can choose whatever priors one would like, but an updater that increases belief that P after receiving and accepting not-P cannot be a Bayesian updater. The belief disconfirmation effect's power to break the Bayesian stalemate lies in its perversity: it dictates that one increases their belief when one accepts that the belief is under legitimate threat (11)

Q1: What exactly is the claim here? If P = *Jesus is the son of God*, Mandelbaum interprets those who said "they believe the article" have accepted $\neg P$. Is that right?

Why doesn't it mean they have accepted (something like) *this is a real article, the World Council really did say that, etc...*? It seems clear that they *don't* believe $\neg P$. Unless he thinks they have inconsistent beliefs? Self-deception?

Q2: Is this qualitatively different than the Kelly-style biased assimilation results? Suppose I accept the article as true and then explain it away ("This is probably God testing my faith"); couldn't that lead to an increase in belief?

Q3: How widespread is this effect? The small sample and sparse replication (most of what he cites is standardly taken to be instances of biased-assimilation) provide reasons for doubt.

V. The Psychological Immune System

Claims that this is not a performance deviation from a Bayesian norm; rather, this i the belief system *functionally properly*.

Cognitive-Dissonance picture (Festinger 1954): receiving evidence that disconfirms a belief which you strongly identify with leads to a negative, phenomenologically distinctive state of "cognitive dissonance". The resulting drive to change beliefs is spurred by a desire to relieve this dissonant feeling, rather than to get to the truth.

Proper functioning: maintaining self-image and motivation. So this is a different *computational-level* analysis, right?

Predicts Batson results, since only those who believed the story were put into a dissonant state, so only they needed to ramp up their belief.

Q1: How exactly does this work? If still have evidence from study, shouldn't increasing their belief give rise to *more* dissonance?

Q2: How adjudicate between this and idea that negative affect is the body/brain's signal that motivates further evidence processing?

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.

Festinger, Leon, 1954. 'A theory of social comparison processes'. *Human relations*, 7(2):117–140.

Jern, Alan, Chang, Kai Min K., and Kemp, Charles, 2014. 'Belief polarization is not always irrational'. *Psychological Review*, 121(2):206–224.

Kelly, Thomas, 2008. 'Disagreement, Dogmatism, and Belief Polarization'. *The Journal of Philosophy*, 105(10):611–633.

Mandelbaum, Eric, 2018. 'Troubles with Bayesianism: An introduction to the psychological immune system'. *Mind & Language*, 1–17.

Marr, David, 1982. *Vision*. MIT Press.