Can resources save rationality? "Anti-Bayesian" updating in cognition and perception*

Eric Mandelbaum[1] (emandelbaum@gc.cuny.edu), Isabel Won[2] (iwon1@jhu,.edu), Steven Gross[2] (sgross11@jhu.edu), & Chaz Firestone[2] (chaz@jhu.edu)

[1] Baruch College, CUNY [2] Johns Hopkins University

Resource rationality may explain suboptimal patterns of reasoning; but what of "anti-Bayesian" effects where the mind updates in a direction *opposite* the one it should? We present two phenomena — belief polarization and the size-weight illusion — that are not obviously explained by performance- or resource-based constraints, nor by the authors' brief discussion of reference repulsion. Can resource rationality accommodate them?

Resource rationality takes seemingly irrational behaviors and reframes them as rational or optimal given other constraints on agents. For example, anchoring-and-adjustment and overestimating extreme events turn out be "rational" after all, by reflecting the rational *allocation* of cognitive resources. Thus, even for such classically irrational phenomena, "the resulting train of thought eventually converges to the Bayes-optimal inference" (p.38).

In such cases, reasoners *fall short* of perfectly rational updating, and it is illuminating that resource- and performance-based constraints can accommodate such suboptimal reasoning. But what about cases where we behave not merely suboptimally, but rather *against* the norms of Bayesian inference? Here, we explore cases where the mind is moved by prior knowledge in precisely the *reverse* direction of what a rational analysis would recommend. These cases are not merely suboptimal, but rather "*anti*-Bayesian", for actively defying Bayesian norms of inference. We consider two such phenomena: belief polarization and sensory integration. Can resource rationality handle them?

First, belief polarization. Receiving evidence contrary to your beliefs should soften those beliefs, even if ever-so-slightly. But this isn't what actually happens when the beliefs in question are central to one's identity — in belief polarization, contrary or disconfirming evidence causes more *extreme* beliefs, not more moderate ones. A classic example was vividly documented by Festinger, Riecken, and Schachter (1956): Cult members who predict the world will end on some date — but who then see that date come and go with no cataclysm — end up *strengthening* their beliefs in the cult's tenets, not softening them. In other words, credible evidence *against* their worldview only makes them hold that worldview more strongly — directly defying Bayesian inference norms.

The same phenomenon can be found under laboratory conditions. For example, one study exposed people who believe that Jesus is the Son of God to a (fake) news article reporting that archeologists had unearthed carbon-dated letters from the New-Testament authors; the letters said the Bible was fraudulent and that its authors knew Jesus was not divinely born (Batson, 1975). Subjects who did not believe the article's content left their beliefs about Jesus unchanged; but, fascinatingly, subjects who did believe the article's content ended up *strengthening* their belief that Jesus was the Son of God. In other words, *affirming* new evidence *against* Jesus's divine birth (~P) caused stronger beliefs in Jesus's divine birth (P). Similar "backwards" updating is also observed for beliefs about nuclear safety (Plous, 1991), health (Liberman & Chaiken, 1992), and affirmative action and gun control (Taber & Lodge, 2006; see also Mandelbaum, 2018).

Why does this happen? In fact, belief polarization is not so mysterious: It has been known for decades, and it is even a predictable consequence of dissonance theory — "the psychological immune system" (Gilbert et. al, 1998) — applied to one's values. What *is* mysterious is why this should occur *in a Bayesian mind* — even one constrained by "resources". Belief polarization is irrational not because

* **Version: 5/10/19**. This is an invited commentary on Lieder & Griffiths, "Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources". Both the target article and this commentary are forthcoming in *Behavioral and Brain Sciences*; but this commentary (as well as the target article itself) may appear slightly differently in published form than here in this version.

people are *insufficiently moved* by evidence, but rather because people are moved in the direction *opposite* the one they should be. And, importantly, these patterns cannot be explained by biased attitudes toward the evidence's source. For example, Bayesian models of milder forms of belief polarization (e.g., Jern et al., 2014) suggest that subjects infer that contrary evidence must have come from unreliable sources (e.g., biased testimony); but this seems inapplicable to the above cases, where the sources are either nature itself (e.g., the world failing to end), or evidence the subject has actively accepted (e.g., news articles they endorsed).

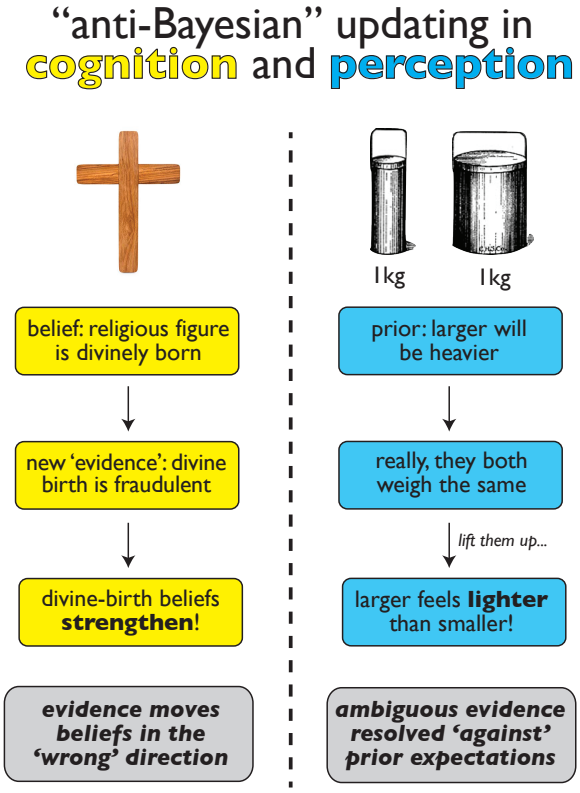

Figure 1. Examples of "anti-Bayesian" updating in the mind. (**a**) Under conditions of cognitive dissonance, acquiring – and affirming – evidence against one's beliefs can cause those beliefs to strengthen (Batson, 1975), whereas Bayesian norms of inference recommend softening those beliefs. (**b**) In the size-weight illusion, one is shown two objects of different sizes but equal weights; when one lifts them up, the smaller one feels illusorily heavier than the larger one (Buckingham, 2014; Charpentier, 1891; Won et al., 2019). In other words, ambiguous sensory data about which of two objects is heavier is resolved "against" one's prior expectations, rather than in favor of one's priors as recommended by Bayesian norms of inference. Can resource rationality accommodate such paradigmatically "irrational" phenomena?

Indeed, "anti-Bayesian" updating is widespread, occurring even in basic perceptual processes. When we have prior expectations about new and uncertain sensory data, rational norms of inference say we should interpret such data with respect for those priors; "people should leverage their prior knowledge about the statistics of the world to resolve perceptual uncertainty" (p.40). But sensory integration frequently occurs the opposite way. Consider the size-weight illusion, wherein subjects see two equally weighted objects — one large and one small — and then lift them both to feel their weight. Which feels heavier? We "should" resolve the ambiguous haptic evidence about which object is heavier *in favor* of our priors; but instead, the classic and much-replicated finding is that we experience the smaller object as *heavier* than the equally-weighted larger object (Buckingham, 2014; Charpentier, 1891). This too is "irrational" — not for *falling short* of Bayesian norms of inference, but for proceeding opposite to them, since we resolve the ambiguous sensory evidence — two equally weighted objects — *against* the larger-is-heavier prior, not in favor of it (Brayanov & Smith, 2010; Buckingham & Goodale, 2013). Indeed, this backwards pattern of updating is so strong that it can produce outcomes that are not merely odd or improbable, but even "impossible" (Won, Gross, & Firestone, 2019): If subjects are shown three boxes in a stack — Boxes A, B, and C — such that Box A is heavy (250g) but Boxes B&C are light (30g), then subjects who lift Box A alone and then Boxes A+B+C together report that Box A feels heavier than Boxes A+B+C — an "impossible" experience of weight (since a group could never weigh less than a *member of that group*).

How can a "rational" account — even a resource-rational one — explain this? Lieder and Griffiths accommodate other sensory "repulsion" effects (Wei & Stocker, 2015, 2017), but that modeling work seems inapplicable to the size-weight illusion. And whereas the original size-weight illusion could perhaps have a tortuous Bayesian explanation (Peters et al., 2016), Won et al.'s modification seemingly cannot: First, it's unclear if previous models of simultaneous lifting apply to Won et al.'s temporally-extended case; but second, there is just no logical chain of reasoning that should end with A alone being heavier than A+B+C together.

More generally: What are the principles that lead to perverse "anti-Bayesian" updating? Perhaps resource rationality wasn't intended to cover all cases (in which case it is not an "Imperial Bayesian" theory; Mandelbaum 2018). But the problem isn't merely that there are counterexamples to resource rationality, but rather that these are predictable, law-like counterexamples that do not reflect performance constraints between interacting mental processes. Indeed, when it comes to these more entrenched patterns, even "resources" may not save rationality.

References

Batson, C. D. (1975). Rational processing or rationalization? The effect of disconfirming information on a stated religious belief. *Journal of Personality and Social Psychology, 32*, 176–184.

Brayanov, J. B., & Smith, M. A. (2010). Bayesian and "anti-Bayesian" biases in sensory integration for action and perception in the size–weight illusion. *Journal of Neurophysiology, 103*, 1518–1531.

Buckingham, G. (2014). Getting a grip on heaviness perception: a review of weight illusions and their probable causes. *Experimental Brain Research, 232*, 1623–1629.

Buckingham, G., & Goodale, M. A. (2013). When the predictive brain gets it really wrong. *Behavioral and Brain Sciences, 36*, 208–209.

Charpentier, A. (1891). Analyse experimentale: De quelques elements de la sensation de poids. [Experimental analysis: On some of the elements of sensations of weight]. *Archives de Physiologie Normale et Pathologique, 3*, 122–135.

Festinger, L., Riecken, H. W., & Schachter, S. (1956). *When prophecy fails*. Minneapolis, MN: University of Minnesota Press.

Gilbert, D. T., Pinel, E. C., Wilson, T. D., Blumberg, S. J., & Wheatley, T. P. (1998). Immune neglect: a source of durability bias in affective forecasting. *Journal of Personality and Social Psychology*, 75, 617–638.

Jern, A., Chang, K.-M. K., & Kemp, C. (2014). Belief polarization is not always irrational. *Psychological Review*, 121, 206–224.

Liberman, A. & Chaiken, S. (1992). Defensive processing of personally relevant health messages. *Personality and Social Psychology Bulletin, 18*, 669–679.

Mandelbaum, E. (2018). Troubles with Bayesianism: An introduction to the psychological immune system. *Mind & Language*. Published online ahead of print.

Peters, M. A. K., Ma, W. J., & Shams, L. (2016). The Size-Weight Illusion is not anti-Bayesian after all: A unifying Bayesian account. *PeerJ, 4*, e2124.

Plous, S. (1991). Biases in the assimilation of technological breakdowns: Do accidents make us safer? *Journal of Applied Social Psychology, 21*, 1058–1082.

Taber, C. S. & Lodge, M. (2006). Motivated skepticism in the evaluation of political beliefs. *American Journal of Political Science, 50*, 755–769.

Wei, X.-X., & Stocker, A. A. (2015). A Bayesian observer model constrained by efficient coding can explain 'anti-Bayesian' percepts. *Nature Neuroscience, 18*, 1509–1517.

Wei, X.-X., & Stocker, A. A. (2017). Lawful relation between perceptual bias and discriminability. *Proceedings of the National Academy of Sciences, 114*, 10244–10249.

Won, I., Gross, S., & Firestone, C. (2019). Impossible somatosensation. *PsyArXiv*.