Whittlestone 2017, Ch. 2: The evidence for confirmation bias

Kevin Dorst
kevindorst@pitt.edu

Rationality Seminar
March 15, 2021

I. The Hurdles

Whittlestone argues that in order to establish that an empirical tendency is an instance of "confirmation bias" (in the pejorative sense), three things must be done:

1) The empirical tendency must be robust/systematic.
2) The tendency must robustly lead to confirmation (or maintenance of) people's prior beliefs.
3) The way in which it does so must violate a clear [and correct!] normative standard.

She finds, across a large literature, that few or none of the classes of research into confirmation bias clear all three hurdles.

This includes:

- Hypothesis testing
- Selective exposure
- Myside bias
- Bias in search time
- Pseudodiagnosticity
- Biased assimilation
- Belief persistence
- Persistence of misinformation and "backfire effect"
- Conservatism
- "Overconfidence"

Let's take a look!

II. The Findings

In each case, we'll consider (1)–(3).

Hypothesis Testing:

Finding: people tend to exhibit a "positive test strategy" (PTS) when testing hypotheses, meaning that if H is their hypothesis, they tend to ask questions 'q'? such that $P(q|H)$ is high—they expect the answer to be 'yes' if their hypothesis is true.

Wason 1960, 1966

Robust? Not clearly. In many "ecological" settings, people seem to be less likely to use PTS (e.g. "if someone's drinking, they're over 21"), and seem to be guided by diagnosticity of the question.

I.e. degree to which answer would favor hypothesis over alternatives: $P(q|H)$ is different than $P(q|\neg H)$.

Confirmatory? Not clearly. Fails to identify false negatives, but good at identifying false positives. So long as you update like a Bayesian,

In triplet-task, cases that don't fall under your hypothesis but are permitted. Not sure how easy this notion is to generalize.

is not expected to raise credence in hypothesis.

Non-normative? Again, not clearly. If what questions you should ask are determined by expected information gain [or expected accuracy], it depends on your priors! For example, consider this variant on Wason selection (card) task. Which cards to flip to test whether all ravens are black? (Objects on one side; colors on other.)

Oaksford and Chater 1994, 2003

Raven	Shoe	White	Black

'White' might refute, but 'Black' might be confirming instance. If $P(raven|white)$ is very low, and $P(raven|black)$ is not as low, then best two options can be 'Raven' and 'Black'.

Selective Exposure

Finding: People have a tendency to look more at congenial sources of information than uncongenial ones.

Frey 1986; Hart et al. 2009

Robust? Very mixed finding, with many "moderators".

Confirmatory? Depends on how you engage! If you scrutinize everything you look at, looking at congenial sources is more likely to *lower* your credence; looking at uncongenial ones is more likely to *raise* it.

Non-normative? No clear normative standards for when and why you should do an "unbiased" search, i.e. look equally at congenial and uncongenial sources.

E.g. if how much you trust a source (how reliable you expect it to be) is correlated with congeniality, then more expectedly accurate to look at congenial sources.

Myside Bias

Finding: The tendency to selectively search memory for congenial information and arguments.

Note: "myside bias" gets used in a lot of different ways. Mercier 2017

Robust? May be (Perkins et al. 1986; Toplak and Stanovich 2003), though Whittlestone says there's not a ton of data within this "search for arguments" paradigm.

Confirmatory? Some evidence that this is driven by focus on a single salient hypothesis—e.g. people do this even when constructing arguments against their own views (Wolfe and Britt 2008).

So perhaps a tendency to construct one-sided arguments.

Non-normative? Depends crucially on goals of the task! Which are often not very clear. If you think good arguments focus only on reasons for one side, and job is to construct a good argument, then makes sense to do this.

[Q: how do we know their *search* is biased? All we see is results of search, presumably. If they know more arguments for one side (or if those are easier to retrieve, since one can lead to the other), the search process itself may be unbiased.]

Think of a random walk. Prior equally likely to end up anywhere, but once you start in one direction, more likely to end up on one side.

Bias in search time:

Finding: People tend to continue searching for new information more often when what they have doesn't fit with what they want to think, and tend to keep searching less when it does.

May be driven by selective scrutiny. Ditto and Lopez 1992

Robust? Not many studies focusing on *confirmation* in particular; more in motivated-reasoning literature.

Confirmatory? Seems to be hinged on motivations; only confirmatory if motivation is to believe what you currently believe.

Non-normative? Not a clear normative standard for information search applied here.

Note: Kelly 2008 is going to argue that selective scrutiny *is* epistemically reasonable, and so if that's the mechanism, this may be caused by rational mechanisms.

Pseudodiagnosticity

Finding: People focus on finding information that is likely when their preferred hypothesis is true ($P(q|H)$ is high), rather than information that is *more* likely if their hypothesis is true than false ($P(q|H)$ is high*er* than $P(q|\neg H)$) (Fischhoff and Beyth-Marom 1983).

E.g. ask if patient has a cough, which would be likely both if have the flu (H), *or* if have some non-flu respiratory illness ($\neg H$).

Robust? Unclear. It varies with the case, and seems to depend on what people believe about the likelihoods. E.g. in versions where can ask about different likelihood ratios, if know $P(q_1|H)$, asking for $P(q_2|H)$ rather than $P(q_1|\neg H)$ seems correlated with whether already have firm opinions about the latter. Often diagnosticity *is* a good guide to what questions people ask (Trope and Bassok 1983).

E.g. q_1 = car can go 60 mph (ask for $P(q_2|H)$), versus q_1 = car can go 120 mph (ask for $P(q_1|\neg H)$) (Feeney et al. 2008).

[**Q:** In what sense could this be a description of what people are doing? People don't ask "Does 2+2=4?", despite the fact that $P(2+2=4|H) = 1$.]

Confirmatory? Only if people aren't updating in a Bayesian way, and accounting for the likelihood of getting various answers given the question they asked (Klayman 1995; McKenzie 2004).

Non-normative? Unclear. Diagnosticity is a clearer normative standard than some others, but it can compete with others.

Biased assimilation

Finding: People tend to interpret mixed evidence as favoring their prior beliefs.

Lord et al. 1979; Taber and Lodge 2006

Robust? Seemingly so, though some replication failures (Kuhn and Lao 1996).

Confirmatory? Yes, seemingly.

Non-normative? Wait for the Kelly 2008 / McWilliams 2019 back and forth next week! Not obviously, since prior beliefs *should* affect how you react to new evidence.

Also Jern et al. 2014

Belief persistence (Ross et al. 1975)

Also persistence of misinformation and "backfire effect" (Nyhan and Reifler 2010)

Finding: People tend to maintain their beliefs after the reasons for them have been debunked. E.g. debriefing method: give subjects feedback that they're reliable at a task, then tell them that the feedback was bunk, and they still are somewhat inclined to think they're reliable.

Or fact-checking false news reports.

Robust? Seemingly so for belief persistence.

Backfire effect has had some major replication failures since (Wood and Porter 2018).

Confirmatory? Yes—at least, in terms of maintaining current belief.

Non-normative? Not clear. Suppose you have some reason to distrust the experimenter—as you now do. If shouldn't be sure that feedback was bunk, then should end up thinking you're more reliable than you started.

E.g. Suppose prior is $P(reliable) = 0.5$. Then get feedback such that $P(reliable|feedback) = 0.8$. Then told feedback is bunk. What now?

- Well, $P(reliable|feedback, bunk) = P(reliable) = 0.5$.
- But $P(reliable|feedback, \neg bunk) \geq P(reliable|feedback) = 0.8$
- And presumably whether or not the feedback is bunk *screens off* whether or not they *told* you the feedback is bunk: $P(reliable|feedback,bunk,told\ bunk) = P(reliable|feedback,bunk)$; and $P(reliable|feedback, \neg bunk,told\ bunk) = P(reliable|feedback, \neg bunk)$
- $\Rightarrow$ as long as you don't *fully* trust what they tell you, should remain more confident you're reliable than your started out:

Let r = *reliable*, b = *bunk*, tb = *told bunk*.

$$\begin{aligned} P(r|f, tb) &= P(b|f, tb) \cdot P(r|f, b, tb) + P(\neg b|f,tb) \cdot P(r|f, \neg b, tb) && \text{Total probability} \\ &= P(b|f, tb) \cdot P(r|f, b) + P(\neg b|f,tb) \cdot P(r|f, \neg b) && \text{Screening off} \\ &\geq P(b|f, tb) \cdot 0.5 + P(\neg b|f,tb) \cdot 0.8 \\ &> 0.5 = P(r) && \text{Don't fully trust them, so that } P(bunk|feedback,told\ bunk) < 1. \end{aligned}$$

Conservatism

Finding? In simple bookbag-and-pokerchips setup, people update in way that's more conservative than Bayesian would (Edwards 1982).

Robust? Seemingly so.

Confirmatory? Subtle. Not for pushing opinion in p in a particular direction. [*But* if we think of "your beliefs" as your priors, does seem to skew you toward *them*.]

Non-Normative? Depends on whether people are certain about the setup! If they fully trust the experimenters, then yes; but if not, then conservative response is actually the Bayesian reaction (Corner et al. 2010).

Remember Anderson on underestimating uncertainty!

"Overconfidence" (Lichtenstein et al. 1982; Hoffrage 2004; Moore et al. 2015)

= over*calibration*

Finding: Beliefs tend to be stronger than accuracy would warrant. For example, of the things people are 80% confident in, only 60% are true. Of the intervals that people say they're 98% confident contain the true value, on 60% of them do.

Robust? Yes.

Confirmatory? Unclear. Could be either a consequence of, or a *cause* of confirmation bias. (If very confident, more likely to engage in selective scrutiny, e.g.)

Non-normative? [Calibration is *not* clearly right normative standard in these sorts of scenarios—often Bayesians should not be expected to be calibrated]

References

Corner, Adam, Harris, Adam, and Hahn, Ulrike, 2010. 'Conservatism in belief revision and participant skepticism'. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 32.

Ditto, Peter H and Lopez, David F, 1992. 'Motivated skepticism: Use of differential decision criteria for preferred and nonpreferred conclusions.' *Journal of personality and social psychology*, 63(4):568.

Edwards, Ward, 1982. 'Conservatism in Human Information Processing'. *Judgment under Uncertainty: Heuristics and Biases*, 359–369.

Feeney, Aidan, Evans, Jonathan, and Venn, Simon, 2008. 'Rarity, pseudodiagnosticity and Bayesian reasoning'. *Thinking & Reasoning*, 14(3):209–230.

Fischhoff, Baruch and Beyth-Marom, Ruth, 1983. 'Hypothesis evaluation from a Bayesian perspective.' *Psychological review*, 90(3):239.

Frey, Dieter, 1986. 'Recent Research on Selective Exposure to Information'. *Advances in Experimental Social Psychology*, 19:41–80.

Hart, William, Albarracín, Dolores, Eagly, Alice H, Brechan, Inge, Lindberg, Matthew J, and Merrill, Lisa, 2009. 'Feeling validated versus being correct: a meta-analysis of selective exposure to information.' *Psychological bulletin*, 135(4):555.

Hoffrage, Ulrich, 2004. 'Overconfidence'. In *Cognitive illusions: A handbook on fallacies and biases in thinking, judgement and memory*, 235–254.

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.

Klayman, Joshua, 1995. 'Varieties of confirmation bias'. *Psychology of learning and motivation*, 32:385–418.

Kuhn, Deanna and Lao, Joseph, 1996. 'Effects of Evidence on Atittudes: is Polarization the Norm?' *Psycholohical Science*, 7(2):115–120.

Lichtenstein, Sarah, Fischhoff, Baruch, and Phillips, Lawrence D., 1982. 'Calibration of probabilities: The state of the art to 1980'. In Daniel Kahneman, Paul Slovic, and Amos Tversky, eds., *Judgment under Uncertainty*, 306–334. Cambridge University Press.

Lord, Charles G., Ross, Lee, and Lepper, Mark R., 1979. 'Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence'. *Journal of Personality and Social Psychology*, 37(11):2098–2109.

McKenzie, Craig R M, 2004. 'Framing effects in inference tasksâĂŤand why they are normatively defensible'. *Memory & cognition*, 32(6):874–885.

McWilliams, Emily C., 2019. *Evidentialism and belief polarization*. 8. Springer Netherlands.

Mercier, Hugo, 2017. 'Confirmation bias—Myside bias.' *Cognitive illusions: Intriguing phenomena in thinking, judgment and memory, 2nd ed.*, 99–114.

Moore, Don A, Tenney, Elizabeth R, and Haran, Uriel, 2015. 'Overprecision in judgment'. *The Wiley Blackwell handbook of judgment and decision making*, 2:182–209.

Nyhan, Brendan and Reifler, Jason, 2010. 'When corrections fail: The persistence of political misperceptions'. *Political Behavior*, 32(2):303–330.

Oaksford, Mike and Chater, Nick, 1994. 'A Rational Analysis of the Selection Task as Optimal Data Selection'. *Psychological Review*, 101(4):608–631.

———, 2003. 'Optimal data selection : Revision , review , and reevaluation'. *Psychological Bulletin & Review*, 10(2):289–318.

Perkins, D N, Bushey, Barbara, and Faraday, Michael, 1986. 'Learning to reason'. *Unpublished manuscript, Harvard graduate School of education, Cambridge, MA*.

Ross, Lee, Lepper, Mark R, and Hubbard, Michael, 1975. 'Perseverance in Self-Perception and Social Perception : Biased Attributional Processes in the Debriefing Paradigm'. *Journal of Personality and Social Psychology*, 32(5):880–892.

Taber, Charles S and Lodge, Milton, 2006. 'Motivated Skepticism in the Evaluation of Political Beliefs'. *American Journal of Political Science*, 50(3):755–769.

Toplak, Maggie E and Stanovich, Keith E, 2003. 'Associations between myside bias on an informal reasoning task and amount of postâĂŘsecondary education'. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, 17(7):851–860.

Trope, Yaacov and Bassok, Miriam, 1983. 'Information-gathering strategies in hypothesis-testing'. *Journal of Experimental Social Psychology*, 19(6):560–576.

Wason, Peter C, 1960. 'On the failure to eliminate hypotheses in a conceptual task'. *Quarterly journal of experimental psychology*, 12(3):129–140.

———, 1966. 'Reasoning'.

Whittlestone, Jess, 2017. 'The importance of making assumptions : why confirmation is not necessarily a bias'. (July).

Wolfe, Christopher R and Britt, M Anne, 2008. 'The locus of the myside bias in written argumentation'. *Thinking & reasoning*, 14(1):1–27.

Wood, Thomas and Porter, Ethan, 2018. 'The Elusive Backfire Effect: Mass Attitudes' Steadfast Factual Adherence'. *Political Behavior*, 1–29.