How explanation guides belief change

Igor Douven

How explanation guides belief change

Article *in* Trends in Cognitive Sciences · August 2021

DOI: 10.1016/j.tics.2021.07.009

CITATIONS
3

READS
52

1 author:

Igor Douven
French National Centre for Scientific Research
205 PUBLICATIONS **2,896** CITATIONS

SEE PROFILE

How Explanation Guides Belief Change

Igor Douven
IHPST / CNRS
igor.douven@univ-paris1.fr

Abstract

Philosophers have argued that people ought to change their graded beliefs via Bayes' rule. Recent work in psychology indicates that people sometimes violate that rule by attending to explanatory factors. Results from computational modeling suggest that such violations may actually be rational.

Keywords: Bayes' rule; belief change; explanatory reasoning; inference; New Paradigm psychology of reasoning; probability.

Evidence indicates that people's graded beliefs tend to obey the probability axioms. Less is known about how humans change their graded beliefs. Philosophers have argued that people ought to do so via Bayes' rule. This paper (i) looks at recent evidence from work on higher-level cognitive functioning showing that people sometimes violate that rule by attending to explanatory factors and (ii) cites results from computational modeling suggesting that such violations may actually be rational, in particular that they help one converge to the truth faster.

Rational reasoning

Classical logic was long believed to provide the standards of correct reasoning. More recently, it has become commonplace to hold that we can do justice to the complexities of human reasoning only if we acknowledge that people's beliefs can differ in degree, and that logic offers little guidance as to how people ought to regulate their degrees of belief. Proponents of the so-called New Paradigm in the psychology of reasoning have done much to popularize the idea that probability theory rather than logic embodies the principles of rational reasoning. Over the past 20 years, evidence has accumulated that, by and large, people's degrees of belief do tend to obey the axioms of probability, although problems have also been identified where people tend to trip up and violate those axioms [1].

Time is not a parameter in probability theory, and so this theory is silent about how people ought to change their degrees of belief when they receive new information. Philosophers have argued that people ought to change their degrees of belief in accordance with Bayes' rule. This is to say that, upon learning A, you should adopt as your new unconditional degrees of belief your degrees of belief conditional on A as they were just prior to learning A. For example, if you believe your favorite football team has a 70 percent chance of winning the upcoming match on the supposition that their top players will be fit enough to play, then upon learning that their top players are fit enough to play, you ought to believe to a degree of .7 that your favorite football team will win the match. According to Bayesians, failure to obey this principle indicates irrationality on your part. (See Box 1 for details.)

While philosophers and statisticians have been mainly interested in the normative aspects of the Bayesian proposal, psychologists have also looked at its descriptive adequacy. And much experimental work aimed at testing the proposal has come up with negative results, showing that, often, people's belief changes deviate from what Bayes' rule prescribes, sometimes quite starkly so.

Abduction

There has been a tendency in the literature to regard these findings as evidence of performance failures, akin to the violations of probability theory mentioned above. But researchers have discovered patterns in these supposed failures which appeared to be caused by people giving special weight in their belief changes to explanatory considerations [2, 3, 4, 5]. This discovery dovetailed with work in epistemology as well as in the history and philosophy of science, according to which explanation ought to guide belief change. For instance, textual evidence shows that scientists often cite the explanatory power of their theory—specifically, its ability to explain a certain range of data better than any of its competitors—as compelling grounds for accepting the theory, implicitly suggesting endorsement of a rule that often goes by the name of "Inference to the Best Explanation," or "abduction."

According to the textbook version of this rule, we should accept the hypothesis or theory that best explains the available evidence. But the textbook version is a bit rough. Would we want to accept the best explanation even if, in absolute terms, it is quite unsatisfactory? Would we want to accept it even if it is satisfactory but there is an alternative explanation that is nearly as satisfactory? Probably not. Philosophers have proposed refined versions of abduction, one being that we should infer the best explanation of our evidence if, and only if, that explanation is (i) good enough, and (ii) considerably better than the second-best explanation. There is recent evidence that abduction thus formulated does describe how people reason, at least in certain entirely ordinary contexts [2].

Bayesians would likely argue that while a categorical notion of belief has psychological reality, the more fundamental notion is that of graded belief, of belief that can vary in strength. And in the aforementioned version, abduction appears to pertain only to the categorical notion. However, recent research on abduction has looked at probabilistic versions of abduction, versions which make abduction look like Bayes' rule, except that they assign bonus points for explanatory goodness (see Box 1). In a series of experiments, Douven and Schupbach ([3]) compared such a version with Bayes' rule in terms of descriptive adequacy, finding that their participants' sequential degrees-of-belief changes were significantly more accurately predicted by the abductive rule than by Bayes' rule. In a re-analysis of their data, it was further found that participants were, on average, significantly more accurate the more weight they gave in their degrees-of-belief changes to explanatory considerations [6].

Can abduction be rational?

While explanatory reasoning is not per se incompatible with Bayesianism [7], there is clear evidence that at least sometimes such reasoning does lead people to transgress Bayesian norms. Committed Bayesians may regard this as further evidence that human reasoners are liable to performance failure. However, recent publications warrant a more favorable take on the data, by seeing them as evidence that people tend to change their degrees of belief in a rational manner. As explained in Box 1, the arguments for Bayes' rule boil down to the claim that there may

be costs attached to following a non-Bayesian rule, costs which could be avoided by sticking to Bayes' rule. That—according to Bayesians—is what makes non-Bayesian reasoners irrational. The problem with this argument is that, even granting the claim about costs, it only follows that non-Bayesian reasoners are irrational if there are no compensating benefits to non-Bayesian reasoning. And as shown by various authors, especially for abductive reasoning, there can well be such benefits [8, 9, 10, 11]. Most notably, using computer simulations comparing Bayes' rule with various probabilistic versions of abduction, it was found that the abductive rules tend to lead to a faster average convergence to the truth than Bayes' rule—which can obviously benefit reasoners [8, 9].

There is already a considerable amount of work on the role of explanation in various high-level cognitive processes, including categorization, generalization, understanding, and interpreting language and behavior [12]. The question of how explanation guides belief change is still underexplored and could for instance also benefit from input from developmental psychologists [13]. The foregoing will hopefully convince colleagues that the question is well worth studying.*

References

[1] Oaksford, M. & Chater, N., New paradigms in the psychology of reasoning. *Annual Review of Psychology* 71: 305–330, 2019.

[2] Douven, I. & Mirabile, P., Best, second-best, and good-enough explanations: How they matter to reasoning. *Journal of Experimental Psychology: Language, Memory, and Cognition* 44: 1792–1813, 2018.

[3] Douven, I. & Schupbach, J. N., The role of explanatory considerations in updating. *Cognition* 142: 299–311, 2015.

[4] Lombrozo, T., Explanatory preferences shape learning and inference. *Trends in Cognitive Sciences* 20: 748–759, 2016.

[5] Walker, C. M., Lombrozo, T., Williams, J. J., Rafferty, A. N., & Gopnik, A., Explaining constrains causal learning in childhood. *Child Development* 88: 229–246, 2017.

[6] Douven, I., Explanation, updating, and accuracy. *Journal of Cognitive Psychology* 28: 1004–1012, 2016.

[7] Wojtowicz, Z. & DeDeo, S., From probability to consilience: How explanatory values implement Bayesian reasoning. *Trends in Cognitive Sciences* 24: 981–993, 2020.

[8] Douven, I., Optimizing group learning: An evolutionary computing approach. *Artificial Intelligence* 275: 235–51, 2019.

[9] Douven, I., The ecological rationality of explanatory reasoning. *Studies in History and Philosophy of Science* 79: 1–14, 2020.

[10] Glass, D. H., An evaluation of probabilistic approaches to inference to the best explanation. *International Journal of Approximate Reasoning* 103: 184–194, 2018.

*I am greatly indebted to Shira Elqayam, Mike Oaksford, Christopher von Bülow, and two anonymous referees for valuable comments on previous versions of this paper, as well as to Lindsey Drayton for helpful editorial advice.

[11] Trpin, B. & Pellert, M., Inference to the best explanation in uncertain evidential situations. *British Journal for the Philosophy of Science* 70: 977–1001, 2019.

[12] Jern, A., Derrow-Pinion, A., & Piergiovanni, A. J., A computational framework for understanding the roles of simplicity and rational support in people's behavior explanations. *Cognition* 210: 104606, 2021, https://doi.org/10.1016/j.cognition.2021.104606.

[13] Liquin, E. G. & Lombrozo, T., Explanation-seeking curiosity in childhood. *Current Opinion in Behavioral Sciences* 35: 14–20, 2020.

Box 1: Explanatory reasoning—only costs?

According to the betting concept of probability, the degree to which you believe that there will be people on Mars before 2040 is the price in cents at which you are willing to take either side in a bet that pays \$1 if indeed there will be people on Mars before 2040 and nothing otherwise.

Using this concept, Bayesians argue that any failure of your graded beliefs to accord with probability theory betokens irrationality. The argument is that any such failure exposes you to a Dutch book, which is a set of bets that guarantees you a loss.

Probability theory is silent on how to change graded beliefs. Bayesians proposed Bayes' rule as an answer to that question, but there are similar rules which take explanatory factors into account. Consider this schema:

> Let $\mathcal{H} = \{H_1, \ldots, H_n\}$ be a set of n mutually exclusive and jointly exhaustive hypotheses H_i and let Pr and Pr′ designate a person's degrees-of-belief function before and after learning evidence E, where $\Pr(E) > 0$. Then that person updates her graded beliefs by rule $\mathrm{r}(c)$ precisely if, for all j,
>
> $$\Pr'(H_j) = \frac{\Pr(H_j)\Pr(E \mid H_j) + f(H_j, E, \mathcal{H})}{\sum_{k=1}^{n}\big(\Pr(H_k)\Pr(E \mid H_k) + f(H_k, E, \mathcal{H})\big)},$$
>
> with
>
> $$f(H_j, E, \mathcal{H}) = \begin{cases} c & \text{if } H_j \text{ best explains } E, \\ 0 & \text{otherwise,} \end{cases}$$
>
> for some $c : 0 \leqslant c \leqslant 1$.

The rule $\mathrm{r}(0)$ is Bayes' rule; other instances are probabilistic versions of abduction.

According to Bayesians, someone who changes her graded beliefs by a non-Bayesian rule is again Dutch-bookable: she can be offered bets at different points in time which will all appear fair to her but together ensure a loss. However, even if there are costs attached to non-Bayesian belief change, might there not also be compensating benefits? Bayesians never asked, but recent evidence supports a positive answer.