The Ecological Rationality of Explanatory Reasoning Side-projects: Computational Philosophy View project Philosophy of Language View project The Ecological Rationality of Explanatory Reasoning *

Igor Douven

The Ecological Rationality of Explanatory Reasoning

Article *in* Studies In History and Philosophy of Science Part A · June 2019

DOI: 10.1016/j.shpsa.2019.06.004

CITATIONS
3

READS
274

1 author:

Igor Douven
French National Centre for Scientific Research
197 PUBLICATIONS **2,287** CITATIONS

SEE PROFILE

The Ecological Rationality of Explanatory Reasoning*

Igor Douven†
SND / CNRS / Sorbonne University

Abstract

There is growing evidence that explanatory considerations influence how people change their degrees of belief in light of new information. Recent studies indicate that this influence is systematic and may result from people's following a probabilistic update rule. While formally very similar to Bayes' rule, the rule or rules people appear to follow are different from, and inconsistent with, that better-known update rule. This raises the question of the normative status of those updating procedures. Is the role explanation plays in people's updating their degrees of belief a bias? Or are people right to update on the basis of explanatory considerations, in that this offers benefits that could not be had otherwise? Various philosophers have argued that any reasoning at deviance with Bayesian principles is to be rejected, and so explanatory reasoning, insofar as it deviates from Bayes' rule, can only be fallacious. We challenge this claim by showing how the kind of explanation-based update rules to which people seem to adhere make it easier to strike the best balance between being fast learners and being accurate learners. Borrowing from the literature on ecological rationality, we argue that what counts as the best balance is intrinsically context-sensitive, and that a main advantage of explanatory update rules is that, unlike Bayes' rule, they have an adjustable parameter which can be fine-tuned per context. The main methodology to be used is agent-based optimization, which also allows us to take an evolutionary perspective on explanatory reasoning.

Keywords: agent-based optimization; belief updating; ecological rationality; explanation; inference; probability; reasoning.

1 Introduction

There is accumulated evidence that explanation is central to cognition. Studies have shown that it plays a pivotal role in processes such as generalization (Sloman 1997), categorization (e.g., Williams & Lombrozo 2010), language interpretation (e.g., Hobbs 2004; Bunt & Black 2000; Douven 2016a; Douven et al. 2018, 2019), learning and concept acquisition (Sidney, Hattikudur, & Alibali 2015; Rittle-Johnson & Loehr 2017), understanding (Keil 2006), and reasoning (Sloman 1994; Koslowski et al. 2008; Johnston et al. 2017; Douven & Mirabile 2018). There is also evidence that explanation is intimately connected to how people "update" (i.e., revise) their degrees of belief upon the receipt of new information (Pennington & Hastie 1988, 1992, 1993; Hastie & Pennington 2000; Bes et al. 2012). Recent studies indicate that this connection is systematic and that it may result from people's following some probabilistic update rule (Douven & Schupbach 2015a, 2015b). While formally very similar to Bayes' rule, the explanation-based rule or rules people appear to rely on, at least in some contexts, are different from, and in general inconsistent with, that better-known update rule.

*The Supplementary Materials, consisting of the Jupyter notebooks referred to in the paper, can be downloaded from this OSF repository: https://osf.io/uah95/?view_only=79fb50954394444a8947f222295c3d92.

†Address: 1, rue Victor Cousin, 75005 Paris, France, igor.douven@sorbonne-universite.fr.

This raises the question of the normative status of people's updating practices. Might the role explanation plays in updating be a bias, against which we might want to warn people, much in the way we warn students against the base rate fallacy (Hammerton 1973; Eddy 1982)? Or might people be right to update their degrees of belief on the basis of explanatory considerations, in that this offers benefits, whether practical or cognitive, that could not be had otherwise? Various philosophers have argued that any form of reasoning at deviance with Bayesian principles is to be rejected—and so explanatory reasoning, insofar as it flouts Bayes' rule—can only be fallacious (van Fraassen 1989; Rosenkrantz 1992; Ladyman et al. 1997; Joyce 1998; Pettigrew 2016). Besides, there is experimental work showing that the influence of explanation on cognition is not always for the good, for instance, in that it leads us to overestimate probabilities (Koehler 1991) or lets us take into account base rates that are evidentially irrelevant (Johnson, Rajeev-Kumar, & Keil 2016).

This paper argues that it would be wrong to reject explanatory reasoning generally as being irrational. Specifically, it is argued that the kind of explanation-based update rules people seem to adhere to allow them to strike the best balance between being fast learners and being accurate learners. Previous work had compared Bayes' rule and explanation-based update rules along the dimensions of speed and accuracy, finding that while the latter tended to take us to the truth fastest, Bayes' rule was on average somewhat more accurate (see below for details). However, the comparison was made in a purely abstract way, without taking into consideration the contexts in which we might want to use those rules (Douven 2013, 2017; Douven & Wenmackers 2017; Trpin & Pellert 2019). Borrowing from the literature on ecological rationality and related notions (e.g., Gigerenzer & Goldstein 1996; Elqayam 2011, 2012; Todd & Gigerenzer 2012; Arkes, Gigerenzer, & Hertwig 2016), we argue that what counts as the best balance is intrinsically context-sensitive and therefore not a question that can be settled in the abstract. As will be seen, a chief advantage of explanation-based update rules is that, in contrast to Bayes' rule, they have an adjustable parameter which can be fine-tuned per context.

Section 2 reviews recent empirical research on how people change their degrees of belief after receiving new evidence. Results from this research will serve to motivate some of the update rules to be contrasted with Bayes' rule. Section 3 reassesses the main arguments for thinking that deviations from Bayes' rule necessarily betoken irrationality and maintains that these arguments rest on questionable premises. It also questions the universalist conception of rationality that underlies those arguments by pointing at alternative conceptions, arguably more useful and more realistic, according to which cognitive practices are to be judged by their success, operationalized as the extent to which they help us achieve our goals. We use computer simulations implementing an agent-based optimization technique to compare Bayes' rule with various probabilistic forms of explanatory reasoning. This methodology also allows us to take an evolutionary perspective on explanatory reasoning, and thereby to address the question of how people may have come to rely on that type of reasoning. Section 4 describes the simulations and analyzes their main results.

2 The psychology of explanatory reasoning

At the center of much modern thinking about the rational updating of degrees of belief is Bayes' rule. Let $\Pr_{t_1}(\cdot)$ and $\Pr_{t_2}(\cdot)$ designate an agent's degrees-of-belief functions at times t_1 and t_2, respectively, where between t_1 and t_2 the agent receives the information A (and no other information). Then, provided $\Pr_{t_1}(A) > 0$, Bayes' rule dictates that the equality $\Pr_{t_2}(B) = \Pr_{t_1}(B \mid A)$ hold for all propositions B expressible in the agent's language. Bayesians conceive of this rule first and foremost as a normative principle, but at the same time they have been happy to point at experiments in which people complied to the rule, at least by

and large (e.g., Oaksford & Chater 2007). It is fair to say, however, that there is also evidence showing that, at times, people update their degrees of belief in ways starkly *deviating* from Bayesian prescriptions (e.g., Phillips & Edwards 1966; Edwards 1968; Marks & Clarkson 1972; Fischhoff & Lichtenstein 1978; Schum & Martin 1982).

Here, we are not interested in reported violations of Bayes' rule per se, but rather in those that might be evidence of explanatory reasoning. Early instances of such evidence are to be found in experimental work on juror decision making by Pennington and Hastie (1988, 1992, 1993), which showed the importance of the order in which evidence is presented to jurors. Pennington and Hastie's participants were significantly more inclined to judge a defendant guilty when the evidence was presented in an order that facilitated the mental construction of an explanatory story of the crime. They also found that the impact of the different pieces of evidence on their participants' degrees of belief strongly violated Bayes' rule.

More explicitly concerned with comparing Bayesian and explanatory reasoning, Bes et al. (2012) demonstrated that when participants were given information about causal relations among three random variables, alongside explicitly provided correlations among those variables, they based their probability judgments on the causal information only, ignoring the statistical information. These participants were thereby violating the so-called Principal Principle—widely endorsed in the Bayesian community—according to which one's *degree of belief* in a hypothesis, given that the *statistical probability* of that hypothesis being true equals x, should equal x (Lewis 1980). For instance, if it is given that a coin is fair, then our degree of belief that the next flip with this coin will be heads should equal .5.

Bes and colleagues attribute the precise effect that the causal information had on their participants' probability judgments to the ease with which that information could be processed into an explanatory story. They did not ask their participants for judgments of explanatory goodness of the statements whose probabilities these participants were to estimate. Such judgments could have been illuminating and might well have revealed a strong correlation between their participants' probability judgments and their explanation-quality judgments.

This is not too speculative and seems in fact likely in view of empirical work reported in Douven and Schupbach (2015a). That work concerned a sequential probabilistic updating task and tried to determine the degree to which updating was influenced by explanatory factors. More specifically, it involved an experiment in which participants were tested individually, in an experimental paradigm that had been used in some of the earliest studies on updating degrees of belief (e.g., in Phillips and Edwards' study cited above), with one important additional element. At the start of the experiment, participants were individually shown two urns, each with 40 balls in it, one ("urn A") containing 30 black balls and 10 white balls, the other ("urn B") containing 15 black balls and 25 white ones. Participants were informed about these contents, and during the experiment they could consult a pictorial representation of the contents whenever they wished. Then the experimenter tossed a fair coin and, depending on the outcome of the toss, chose either urn A or urn B. Everything was transparent to the participants except for which urn had been selected, which happened outside the participants' sight. Next, the experimenter drew 10 balls from the selected urn, one by one, lining them up before the participants. After each draw, participants were asked three questions, namely: (i) how well the hypothesis that urn A had been selected explained the outcomes of the drawings so far; (ii) the same question, but now regarding the hypothesis that urn B had been selected; and (iii) how likely they thought it was that urn A had been selected, in view of the drawings so far. The questions about explanatory goodness had to be answered by indicating a point on a continuous scale with anchors "extremely poor explanation" and "extremely good explanation." Questions (i) and (ii) had not been asked in previous experimental work using this "bookbag-and-pokerchips" paradigm (Edwards 1968).

In their analysis, Douven and Schupbach fitted three mixed-effects models, all of which had the collected participants' responses to the third question as fixed effect, participants

as random effects, and at least the objective conditional probabilities as predictor variable. In one model, objective conditional probabilities were the only predictor. A second model further included both the collected responses to the first question and the collected responses to the second question as predictors. The third model had as predictor, next to the objective conditional probabilities, the computed *difference* between the participants' responses to the first question and their responses to the second question. The third model did best across all standard comparative tests, followed by the model that included judgments of explanatory goodness as predictors. Most importantly, both these models were *far* superior to the model that tried to predict participants' responses strictly on the basis of objective probabilities.

As was explained in Douven and Schupbach (2015a), these model comparisons carry bad news for Bayesians, at least for those who were hoping that Bayesian norms would also achieve greater predictive accuracy than competing norms. For again by the Principal Principle, from a Bayesian perspective the participants in Douven and Schupbach's experiment should, after each update, have set their degree of belief that urn A had been selected equal to whatever the objective probability was that that urn had been selected, given the registered draws. Most importantly, although some Bayesians (e.g., Lipton 2004; Weisberg 2009) have been happy to admit that explanatory considerations can factor into the determination of prior probabilities and / or likelihoods (though not in the kind of case at hand, where priors and likelihoods are objectively given), there is nothing in Bayesianism that could account for the help that the participants' judgments of explanatory goodness offered in predicting their degrees of belief. In short, to vindicate Bayesianism as a descriptive theory of human updating, the model with only objective conditional probabilities as predictor should have come out on top—which did not happen.

It is consistent with everything said in Douven and Schupbach (2015a) that *how* explanatory considerations influence belief change is unsystematic at least to the extent that it cannot easily be conceived as the result of rule-following behavior, in particular the kind of rules that have surfaced in the philosophical literature under the name of "Inference to the Best Explanation" (e.g., Harman 1965; Boyd 1990; Psillos 1999; Douven 2002; Lipton 2004; Schurz 2008a; McCain & Poston 2014; Poston 2014; Schupbach 2017; Williamson 2018). However, results from Douven and Schupbach (2015b) suggest that people do in fact respond to the receipt of new evidence much as they would if they followed an update rule close to some that philosophers have discussed in the debate about the normative status of explanatory reasoning.

In that paper, Douven and Schupbach used the objective probabilities from their earlier study as well as some probabilistic measures of explanatory goodness to compute, for each participant and each draw separately, the explanatory goodness of both hypotheses at stake in the study (that urn A had been selected, and that urn B had been selected). They then tried to predict the participants' updates again, but now using objective conditional probabilities and computed explanatory goodness of the two hypotheses as predictors. So they basically repeated the analysis summarized above, but now with the *computed* explanatory goodness values in place of the *subjective* judgments of explanatory goodness that had served as predictors in the earlier analysis.

Adding the computed explanatory goodness values as predictors to the model with only objective conditional probabilities yielded a significantly better model. This was true in particular for two measures of explanatory goodness, to wit, Popper's (1959) measure, according to which hypothesis H's power to explain evidence E is given by

$$\frac{\Pr(E \mid H) - \Pr(E)}{\Pr(E \mid H) + \Pr(E)},$$

and Good's (1960) measure, according to which H's power to explain E equals[1]

$$\ln\left(\frac{\Pr(E \mid H)}{\Pr(E)}\right).$$

These measures can be used to define probabilistic versions of explanatory reasoning, as will be seen later on. The results from Douven and Schupbach (2015b) are an indication that their participants were updating, at least implicitly and approximately, according to one of those rules.[2,3]

3 Is explanatory reasoning irrational?

Hard-nosed Bayesians will be unfazed by data showing that people do not always obey Bayes' rule and sometimes base their probability updates on judgments of explanatory goodness. For them, that will be just another addition to the long list of probabilistic biases known in the literature. And it is a bias because people ought to follow Bayes' rule, lest they qualify as irrational. This section looks at the two main arguments Bayesians have propounded for that claim and finds that they rest on feeble grounds. In addition, we look at recent literature on rationality urging a closer tie between accounts of rationality and psychological theories of cognition. This literature suggests that the general Bayesian conception of rationality as comprising a body of universally valid principles is fundamentally misguided.

3.1 The dynamic Dutch book argument

For many years, the claim that Bayes' rule is the only rational update rule was backed by reference to the so-called dynamic Dutch book argument (Teller 1973; Lewis 1999). At the core of this argument is a mathematical result, often referred to as "the dynamic Dutch book theorem," showing that someone updates her degrees of belief in a way at variance with Bayes' rule if, and only if, she is vulnerable to a dynamic Dutch book, meaning a collection of bets that seem individually fair to the updater at the time they are offered, but that jointly guarantee a negative net payoff. The dynamic Dutch book *argument*, then, is that surely no rational person would want to be exposed to this risk. And given that the theorem is a priori, we can all figure out how to avoid the risk, to wit, by complying with Bayes' rule.

It is first to be noted that the mathematical result is perhaps not quite as solid as the epithet "theorem" suggests. Various authors have shown that if we think of update rules not in isolation but rather as being part of a package of principles which also includes decision-theoretic principles (such as Maher's 1993 principle to look ahead before engaging in any bets), then there turn out to be packages that let one update in a non-Bayesian way while avoiding dynamic-Dutch-book vulnerability (Douven 1999; Tregear 2004; Eberhardt & Danks 2011).

From a strategic point of view, the more important criticism of the dynamic Dutch book argument is the one coming from within the Bayesian community. Bayesians grew wary of

[1] Douven and Schupbach (2015b) in fact used a rescaled version of Good's measure; the mathematical details need not detain us here.

[2] Costello and Watts (2016, 2018) try to explain away a number of known deviations from Bayesian norms as being due, at bottom, to memory limitations. As shown in Douven (2019a), however, their strategy is unable to handle the deviations from Bayesian updating documented in Douven and Schupbach (2015a).

[3] As an anonymous referee observed, the experimental results cited in this section all concern explanatory reasoning by laypeople, and it is certainly possible that those results would have been significantly different had the participants been scientists. But however interesting it would be to have scientists participate in experiments on explanatory reasoning, it will be vastly more difficult to recruit, say, thirty scientists for a psychology experiment than it is to recruit (if need be) hundreds of laypeople via one of the common crowdsourcing platforms (such as Amazon's Mechanical Turk, Figure Eight, or Prolific).

the argument because—they came to believe—it addresses the wrong kind of rationality. When concerned with updating degrees of *belief*, the relevant notion of rationality is that of *epistemic* rationality, not *practical* rationality; and all the dynamic Dutch book argument shows (barring the previous point, so supposing it shows anything at all) is that it can be practically disadvantageous to update in a non-Bayesian fashion.

But here Bayesians may well be more critical of their erstwhile favored argument than they ought to be. For their supposition that we can make a clean split between practical and epistemic rationality has come under a cloud. A number of epistemologists have argued that questions concerning our epistemic attitudes—whether we know or are justified in believing something—cannot be answered independently of our practical interests, which may vary from one person to another and even for the same person from one situation to the next (Fantl & McGrath 2009; Rinard 2017). Apart from such possible conceptual connections between the epistemic and the practical, there is an obvious factual connection: what kind of epistemic projects we can pursue depends on all sorts of practicalities. Whether I am in the position to pursue any research projects at all depends on how well I can organize my life, on whether I can secure a living as a researcher, on the degree to which I succeed in protecting my research time, and so on. It thus seems wrong to dismiss out of hand the dynamic Dutch book argument for "merely" exposing practical liabilities for the non-Bayesian updater.

The real problem with that argument lies elsewhere. In essence, the argument is that updating by means of a non-Bayesian rule can cost you money that would stay in your pocket if you updated by Bayes' rule. But note that many things cost money. By dining out you can, and typically will, lose money, and in a foreseeable manner: waiters tend to present you with a bill not long after you have finished your meal. Still, that the loss can be prevented by fixing your own dinner at home is not a compelling argument against dining out; if all goes well, you get something in return that is *worth the expense*. This may hold for non-Bayesian updating rules as well, and so we should not just look at the possible *costs*, but also at possible *benefits*, of such rules. For some reason, Bayesians have almost completely ignored the question of whether non-Bayesian updating might bring benefits that outweigh the costs.

Some first attempts to show that non-Bayesian update rules actually do carry benefits over Bayes' rule were made by Douven (2013, 2017, 2019b) and Douven and Wenmackers (2017).[4] These authors started by noting that, often, we are not only interested in finding out the truth about some matter, but also in finding out the truth sooner rather than later. For probabilistic settings, the foregoing truism can be phrased in terms of the speed with which we come to assign a high probability to the truth (speed of convergence, for short). To see how various update rules can behave differently in this respect, suppose we are receiving binomial data (i.e., independent and identically distributed binary data), the data coming in one at a time and our task being to estimate the probability of "success" (e.g., the probability that the patient survives, or that the treatment is effective, or that the coin lands heads). More concretely, let the data consist of the outcomes of tosses with the same coin, where it is antecedently known that the bias of the coin (for concreteness, the probability of heads) can take eleven possible values, ranging from 0 (certain to land tails) to 1 (certain to land heads), in increments of .1. A priori, none of these values is more likely than any other to be the true bias, so before any tosses have been observed, it is reasonable to be no more confident that the bias takes one particular value than that it takes another.

For the moment, we compare just two update rules; we look at additional ones later on. One rule is Bayes' rule. To state the other, which we shall label "EXPL," let $\{H_i\}_{1 \leqslant i \leqslant n}$ be a set of mutually exclusive and jointly exhaustive hypotheses and let $\Pr_{t_1}(\cdot)$ and $\Pr_{t_2}(\cdot)$ designate an agent's degrees-of-belief function before and after learning evidence E. Then EXPL is an

[4] Trpin and Pellert (2019) show that the results from the mentioned papers generalize to cases where the evidence is itself uncertain. They effectively compare Jeffrey conditionalization (a generalization of Bayes' rule—see Jeffrey 1965) with an explanation-based counterpart of that rule.

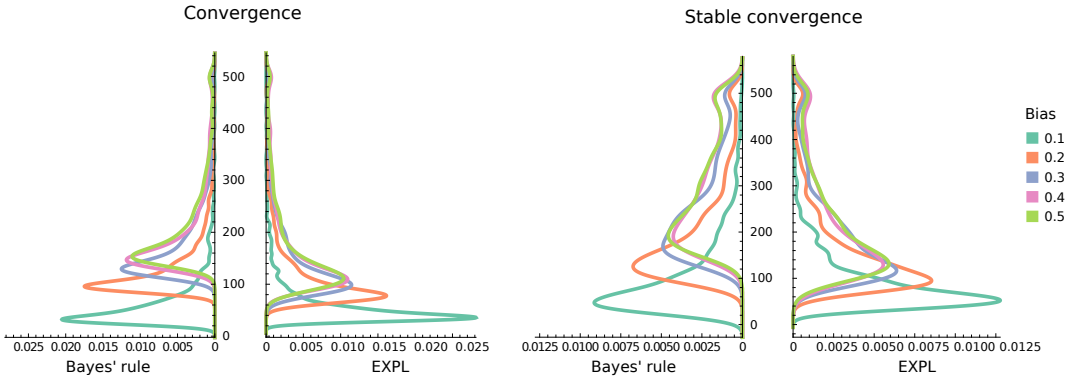

Figure 1: Paired density plots showing speed of convergence to the truth (left) and speed of stable convergence to the truth (right) for Bayes' rule and EXPL, for biases from .1 to .5. The x-axes show probability densities, the y-axes show toss numbers. Individual plots can be interpreted as showing the relative likelihood that the first time a probability above the threshold is assigned (left) / is stably assigned (right) to a hypothesis occurs at a given toss.

instance of this schema:

$$\Pr_{t_2}(H_i) \quad = \quad \frac{\Pr_{t_1}(H_i)\Pr_{t_1}(E \mid H_i) + \mathcal{E}(H_i, E)}{\sum_{j=1}^{n}\left(\Pr_{t_1}(H_j)\Pr_{t_1}(E \mid H_j) + \mathcal{E}(H_j, E)\right)}, \tag{1}$$

where $\mathcal{E}(\cdot,\cdot)$ assigns a bonus c to the hypothesis that explains the evidence best, and nothing to the other hypotheses, and where it is assumed that $\Pr_{t_1}(E) > 0$. EXPL is the instance with $c = 0.05$. (Note that by setting $c = 0$ we obtain Bayes' rule.)

There is considerable leeway in defining explanatory bestness. Here, we adopt a proposal by van Fraassen (1989), who was the first to discuss EXPL as a possible formalization of probabilistic explanatory reasoning, precisely in the context of the kind of statistical model we are considering. Van Fraassen proposed taking a hypothesis about the bias of the coin to be the best explanation of the available evidence at a given time (the outcomes of the tosses up till that time) if that bias was closest to the actually observed frequency of heads in the evidence. For example, if there have been 32 heads in the first 99 tosses, then the hypothesis that the coin has a bias of .3 is the best explanation after the 99$^{\text{th}}$ toss, and so in updating on the outcome of that toss, we should assign the bonus c to the bias = .3 hypothesis. In the case where two biases are equally close to the observed frequency, EXPL splits the bonus c between the corresponding hypotheses.

Douven (2013) compared Bayes' rule and EXPL in the coin tossing setting described above by simulating, for each possible bias, 1,000 sequences of 500 tosses, updating on those tosses via Bayes' rule and via EXPL, respectively, and then registering, for each run, after how many tosses the probability assigned to the true bias (whichever it was for that run) exceeded the threshold value of .9 (called "convergence" here) as well as registering after how many tosses that probability was *stably* above .9 in that it also stayed above .9 for the remaining tosses in the sequence of 500 tosses ("stable convergence"). The simulations were rerun for the present paper, and the results are graphically summarized by the plots shown in Figure 1.[5]

[5]Results for biases from .6 to .9 are not displayed in Figure 1 because they are the mirror images (barring random noise) of the results shown—as they should be, given that a bias of p for heads is equivalent to a bias of $1 - p$ for tails. For the extreme biases, convergence and stable convergence happen at a fixed toss: for a coin that always comes up heads, Bayes' rule assigns a probability to the truth that is above the threshold after 23 tosses, EXPL after 14 tosses. Obviously, convergence and stable convergence coincide in the extreme cases.

The code for the simulations, as well as for the analyses to be reported in this section, is to be found in the Jupyter notebook `Section3.ipynb` in the Supplementary Materials.

It can be seen that EXPL does better than Bayes' rule on both counts. Whether we look at convergence or at stable convergence, the mode of the distribution for EXPL lies around the 100^{th} toss, regardless of the bias of the coin. By contrast, for Bayes' rule the corresponding modes are mostly well above that toss.[6] Of a series of Mann–Whitney U tests, comparing for each bias hypothesis the convergence results from Bayes' rule with those from EXPL, all came out significant (all ps $< .0001$) with the exception of the test for the bias = .1 hypothesis. Exactly the same outcome was obtained for the stable convergence results.

When Bayesians advanced their dynamic Dutch book argument, their main mistake was to focus narrowly on just one specific practical liability of non-Bayesian updating. We should not follow their lead by pretending that speed of convergence, or stable convergence, is all that matters to rational belief change, or all that matters next to Dutch-bookability. While we are better served by an update rule that leads us toward the truth fast, all else being equal, things may *not* be equal, even discounting possible financial losses at the hands of Dutch bookies. Most notably, we have to realize that what makes EXPL fast in terms of convergence to the truth may also give it a greater tendency to lead us astray; that, in other words, the greater speed comes at the expense of accuracy.

To see that this concern is real, note that, with each EXPL update, we can associate a probability that the explanation bonus gets in its entirety assigned to a false hypothesis, and that this probability can be considerable. Suppose, for instance, that in our present model a coin with an actual but to the updater unknown bias of .5 (so a fair coin) is tossed ten times. Then the probability that EXPL will after the tenth toss assign the bonus to the *true* bias hypothesis is the probability that there will be exactly 5 heads in the 10 tosses, which can be calculated to be approximately .25. Consequently, the probability that after ten tosses a *false* hypothesis receives the bonus is about .75.

This probability goes down as the coin is tossed more often, and it is also lower for more heavily biased coins (whether toward heads or toward tails). For example, if the bias were .2 (equivalently, .8) and we were to consider the possibility that the bonus gets assigned to a false hypothesis after the 100^{th} toss, the probability of that actually happening would be about .17; for the 1000^{th} toss, it is smaller than .0001. More generally, the probability of a wrong assignment goes to 0 (regardless of the bias of the coin) as the number of tosses taken into consideration goes to infinity.[7]

But if we are interested in which rule should be used, not by idealized agents, but by us ordinary mortals, then such limit behavior carries much less significance than what happens in the short or medium-long run, in particular how Bayes' rule and EXPL compare in terms of accuracy in the short or medium-long run. One standard way to measure accuracy, as pertaining to degrees of belief, is provided by so-called scoring rules. The historically first such rule, which is still most commonly used, is the Brier score (Brier 1950). Suppose we have n mutually exclusive and jointly exhaustive hypotheses, $H_1, \ldots, H_n$. Then, where $[\![H_i]\!] \in \{0, 1\}$ is the truth value of hypothesis H_i, with 0 designating "false" and 1 designating "true," a person assigning probability $\Pr(H_i)$ to H_i (for each $i \leqslant n$) incurs a Brier penalty of $\sum_{i=1}^{n} ([\![H_i]\!] - \Pr(H_i))^2$. To illustrate, let $\{H_1, H_2, H_3\}$ be a set of mutually exclusive and jointly

[6] This is even clearer in the more detailed (but also more space consuming) violin plots included in the Jupyter notebook mentioned in note 5.

[7] To be more exact, we first note that the normal distribution gives a good approximation to the binomial distribution provided both $np \geqslant 10$ and $n(1-p) \geqslant 10$, where p is the success probability (the bias) and n the number of trials (tosses, in our case). Excluding the extreme bias hypotheses (always heads, always tails), these conditions are satisfied for even moderately large numbers of tosses. Then the probability that the bonus is assigned to a false hypothesis after n tosses of a coin with bias p (where $0 < p < 1$) is approximately

$$1 - \operatorname{erf}\left(\frac{.05}{\sqrt{2}} \cdot \sqrt{\frac{n}{p - p^2}}\right),$$

which converges to 0 as n tends to infinity.

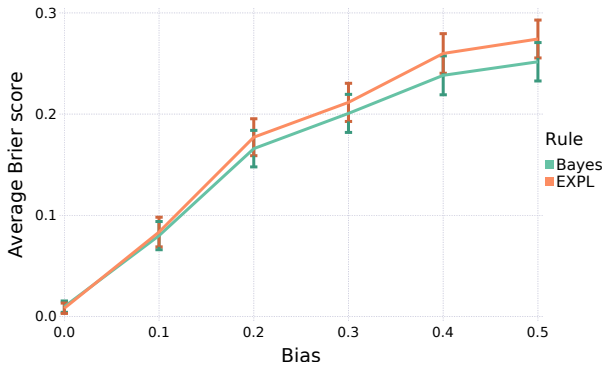

Figure 2: Mean Brier scores over 500 tosses per bias hypothesis, averaged over all 1,000 simulations.

exhaustive hypotheses, which you believe to a degree of .4, .3, and .2, respectively. Then if H_1 happens to be true, you incur a Brier penalty of $(1 - .4)^2 + .3^2 + .2^2 = 0.49$. If instead H_3 had been true, your incurred penalty would have been higher (0.89), which makes sense, given that your degrees of belief would then be intuitively less accurate: in the former case, you are most strongly convinced of the truth; in the latter, you are most strongly convinced of a falsehood and even *least* strongly convinced of the truth. One easily verifies that the minimum Brier penalty of 0 is achieved if, and only if, the truth is believed to the highest degree and no confidence at all is invested in any competing hypotheses.

We can look again at the simulations we ran, calculate for both Bayes' rule and EXPL the Brier penalties they incurred after each update, and then take the average of the penalties for the 500 tosses in the simulation. Averaging once more, now over the 1,000 simulations that were run per bias hypothesis, we obtain the results shown in Figure 2.[8]

For both rules, there is a very strong speed–accuracy trade-off, both when speed is understood as convergence and when it is understood as stable convergence: all rs $> .99$, all ps $< .0001$. Furthermore, as can also be readily seen in Figure 2, EXPL incurs a higher average Brier score than Bayes' rule for most bias hypotheses. However, a series of t-tests on the means per bias hypothesis shows that the difference is never significant: largest $t = 1.6$, smallest $p = .1$.

One might be tempted to see this as a victory for EXPL. Above we saw that EXPL was *significantly* faster in converging toward the truth than Bayes' rule, and now we see that it is *not significantly* less accurate than Bayes' rule. This temptation is to be resisted, however. A main point to be made in Section 4 is that we should not make too much of results like these *in the abstract*: to achieve a proper assessment of update rules, we need to consider them in context, for only then can we make a reasonable judgment as to how accuracy and speed are to be balanced.

3.2 Inaccuracy minimization

There is a newer argument that Bayesians have put forward to show that any deviation from Bayesian updating betokens irrationality (Rosenkrantz 1992; Joyce 1998; Pettigrew 2016). This argument was meant to fill the gap that Bayesians felt had come to exist when they almost collectively rejected the dynamic Dutch book argument as being about the wrong kind of rationality. The new argument purports to provide a defense of Bayes' rule by pointing out

[8] Results are only shown for hypotheses up until bias = .5, for reasons given in note 5.

an explicitly *epistemic* (as opposed to pragmatic) virtue of this rule not shared by other rules. Specifically, the gist of the argument is that unless one updates via Bayes' rule, one's degrees of belief are not as accurate as they could be. In more colloquial terms, if one updates via some non-Bayesian rule, one's mental representation of the world is going to be less sharp than it would be were one to use Bayes' rule instead.

There are at least two reasons to find this argument unsatisfactory. First, its proponents spell out the notion of accuracy in terms of the previously mentioned Brier score. This threatens to undermine the foremost motivation of the argument, however. While Brier (1950) was the first to propose a scoring rule, by now a welter of such rules exists (see, e.g., Cooke 1991; Bickel 2010; Douven 2019c). As Cooke (1991, p. 121) emphasizes, a main question concerning such rules is whether "the score reward[s] those features that we would like subjective probability assessments to display." And it is arguable that which features we would like subjective probabilities to display depends on context, notably also on pragmatic concerns. This will be one of the central points to be made in the next section, but has been argued on independent grounds in Douven (2019c), where it is accordingly concluded that the question of which scoring rule to use cannot be answered without regard to our practical concerns in the context in which the rule is to be used. As a result, the argument from accuracy maximization cannot claim to make a distinctively epistemic case for the unique rationality of Bayes' rule.

Second, however, there is a more fundamental problem with the argument. While it is touted as showing that Bayes' rule maximizes accuracy, looking at the fine print reveals that what the argument really shows is something less grandiose, to wit, that Bayes' rule maximizes *expected* accuracy, and then only *of the next update*. So, judging by your *current* degrees of belief, you expect your *new* degrees of belief to be maximally accurate if you update on your newly acquired evidence via Bayes' rule. But note that the perspective from which you judge this to be the case is precisely the one you are about to abandon, by updating on the new evidence (Williams 2012). Moreover, it is acknowledged by the proponents of the argument that maximizing *expected* accuracy is not an indication that we are maximizing the *actual* accuracy of our soon-to-be-adopted degrees of belief, let alone that we are approaching our longer-term goal of having maximally accurate degrees of belief, of having a maximally sharp representation of the world. In particular, it is left open by the argument, and by anything its proponents have said so far, that non-Bayesian update rules may do a better job in achieving those other goals. Nor has any reason been given to hold that those goals are less worthy of our pursuit.[9]

3.3 Ecological rationality

Whereas *philosophers* identifying as Bayesians are strongly inclined to condemn any violation of Bayesian norms as irrational, the Bayesian community in *psychology* appears to be more open-minded. For example, Oaksford and Chater, who are among the main proponents of Bayesianism in psychology (see, e.g., their 2007), admit that "it is unclear what are the rational probabilistic constraints on dynamic inference" (2013, p. 374). And Elqayam and Evans (2013) are explicit that their brand of Bayesianism only requires our reasoning to be *broadly* probabilistic, meaning in particular that it does not require strict adherence to Bayesian norms

[9] It is in fact ironic that when Berg, Biele, and Gigerenzer (2016) surveyed 125 academics, asking them for their degrees of belief concerning the usefulness of prostate cancer screening, they found literally zero correlation between the degree to which their participants conformed to Bayesian principles and the actual accuracy of their degrees of belief. Even more ironic, they found that the participant who conformed most closely to those principles was also the one whose degrees of belief were most inaccurate. More directly related to explanatory reasoning, a re-analysis of the data from Douven and Schupbach (2015a), discussed in Section 2, showed that participants tended to have more accurate degrees of belief the more weight they had given to explanatory considerations in their updating (Douven 2016b).

(see in the same vein Over 2009 and Elqayam 2018). Hence findings like those discussed in Section 2 may not be too troubling for Bayesian psychologists.

Bayesian philosophers unconcerned with the descriptive adequacy of their position might deem such violations not just not troubling but altogether irrelevant. After all, they can claim to be postulating an ideal to which any rational person should aspire. No amount of data showing that people fail to comply with that ideal would seem to undermine its status *qua* ideal.

Perhaps not, but what does threaten the Bayesian conception of rationality as a whole is that, as various authors have argued, the Bayesian ideal is simply unachievable for limited beings like us (see, e.g., Arkes, Gigerenzer, & Hertwig 2016). Worse still, it is not even clear what steps can be taken to *approach* it (Earman 1992, p. 56). Such concerns have led various researchers to doubt the validity of Bayesianism as a basis for human rationality and to pursue alternatives more closely in touch with psychological theories of cognition.

This alternative approach came to fruition, most prominently, in the development of Simon's (1982) celebrated theory of bounded rationality and Gigerenzer and collaborators' related theory of ecological rationality (e.g., Gigerenzer & Goldstein 1996; Gigerenzer et al. 1999; Gigerenzer 2000; Goldstein & Gigerenzer 2002; Gigerenzer, Hertwig, & Pachur 2011; Todd & Gigerenzer 2012; Todd & Brighton 2016; Schurz & Hertwig 2018). A more recent proposal in this vein is Elqayam's (2011, 2012) account of grounded rationality.

While these accounts differ in their details, they have the important commonality of taking into consideration the various biological and cognitive limitations humans are subject to, as well as the environment or environments in which we, both individually and socially, operate. The said limitations impose constraints on rationality that are to some extent universal—we are *all* finite beings, with finite computational powers, finite memory capacity, and so on—but the environment in which we are to act is not the same for everyone, and may be different for each of us at different points in time; the same holds for the cognitive tools available to us. Especially Gigerenzer and colleagues and also Elqayam argue that, on their accounts, there can be no universally applicable rationality standards that we might be able to pin down a priori. Whether a person's behavior qualifies as rational depends on whether the behavior facilitates achievement of the person's goal or goals, in the context at issue, and given the resources available to the person in that context.

As Elqayam points out, the term "behavior," in the present conception of rationality, is to be understood broadly, so as to include *cognitive* behavior. Concomitantly, the person's goals may be partly or wholly cognitive ones. She thus sees at least her notion of grounded rationality as being in accordance with Evans and Over's (1996) proposal to conceive of instrumental (pragmatic, goal-oriented) rationality as primary and as subsuming epistemic rationality, given that among a person's goals may be ones to obey certain systems of epistemic norms (like classical logic, or the laws of probability). But grounded rationality goes beyond a mere allegiance to instrumental rationality in that its emphasis on the role of context and on individual differences strongly implies that "instrumental rationality cannot be reduced to any one-size-fits-all normative framework" (Elqayam 2012, p. 46; also Elqayam & Evans 2011).

Taking on board this more psychologically oriented approach to rationality, it is even clearer that the dynamic Dutch book argument against non-Bayesian update rules cannot simply be dismissed on the grounds that it deals with the wrong notion of rationality: epistemic rationality *is* a kind of pragmatic rationality. It is then equally clear, however, that a verdict about the rationality (or otherwise) of a given update rule cannot be reached without considering possible users of such rules *as located in a given context, with their specific goals in that context and their specific cognitive capacities.* Only then will it make sense to judge the use of a particular update rule as rational, or as more rational than the use of some alternative rule, and the judgment will have to be based on a cost–benefit analysis of the rule or rules within that context.

Arkes, Gigerenzer, and Hertwig (2016, p. 33) are very specific about how one should investigate the rationality of a cognitive strategy. Their proposal is to begin by identifying (i) the goal of an individual or a group; (ii) the strategies available to the individual or group for achieving that goal; and (iii) the structural properties of the individual's or group's environment. In a second step, one then determines which strategy or strategies are most likely to help the individual or group achieve her / its goal in the given environment. Note how different this procedure is from trying to lay down rationality criteria a priori.

4 Simulating explanatory reasoning

To show that updating via some probabilistic version of explanatory reasoning can be rational, and more rational than updating via Bayes' rule, I will follow the steps from Arkes and colleagues' proposal. Specifically, I will define an environment, agents with a goal in that environment, and a variety of update rules available to the agents, and then use computational simulations to determine which rule or rules is / are most effective in realizing the agents' goal in that environment.

4.1 Setup

Earlier, we compared Bayes' rule with one specific instance of schema (1), one where the explanation bonus c was 0.05; we labeled that instance "EXPL." From now on, I will use this label to refer generically to instances of (1) with $c > 0$, and so to instances that can be said to embody some form of explanatory reasoning. To these, we add the instances of two further schemata, inspired by the finding from Douven and Schupbach (2015b) that Good's and Popper's measures of explanatory goodness (stated at the end of Section 2) were useful in predicting participants' responses from the study reported in Douven and Schupbach (2015a). In Douven (2017), these measures were used to define what were called "Good's rule" and "Popper's rule," respectively. Both rules are instances of this schema:

$$\Pr_{t_2}(H_i) \;=\; \frac{\Pr_{t_1}(H_i)\Pr_{t_1}(E \mid H_i) + c\Pr_{t_1}(H_i)\Pr_{t_1}(E \mid H_i)\mathcal{M}(H_i, E)}{\sum_{j=1}^{n}\left(\Pr_{t_1}(H_j)\Pr_{t_1}(E \mid H_j) + c\Pr_{t_1}(H_j)\Pr_{t_1}(E \mid H_j)\mathcal{M}(H_j, E)\right)}, \tag{2}$$

where $\Pr_{t_1}(\cdot)$ and $\Pr_{t_2}(\cdot)$ are as defined before. Good's rule is obtained from this by substituting Good's measure for $\mathcal{M}(\cdot)$, and Popper's rule by substituting Popper's measure for the same expression. Note that, as stated, these rules are themselves mere schemata from which specific instances are derived by fixing a particular value for $c \in (0, 1]$. This constant c determines what percentage of H_i's probability after a Bayesian update on E is added in proportion to this hypothesis' power to explain E (see Douven 2017 for further explanation). Here, too, we would obtain Bayes' rule by setting $c = 0$.[10]

The rules to be compared in the following—the available strategies—are Bayes' rule, and instances of EXPL, Good's rule, and Popper's rule, where for all these instances $c \in (0, 0.25]$. The setting in which these rules are to be compared is, although stylized, not at all unrealistic. It concerns doctors at an Intensive Care Unit (ICU) who try to determine what exactly is wrong with the patients who are rushed in, and who, based on the information they obtain from the tests that are carried out, must choose an intervention. They are under some pressure to act fast: as time passes, the probability that the patient will die increases, and making the right intervention *decreases* that probability. On the other hand, making the wrong intervention *increases* probability of death.

[10] In Douven (2017), these rules were compared with one another and with Bayes' rule and EXPL in terms of speed and accuracy, in the manner described in Section 3.1 for Bayes' rule and EXPL. They turned out to be faster than Bayes' rule but slower than EXPL, while being more accurate than EXPL but (slightly) less accurate than Bayes' rule.

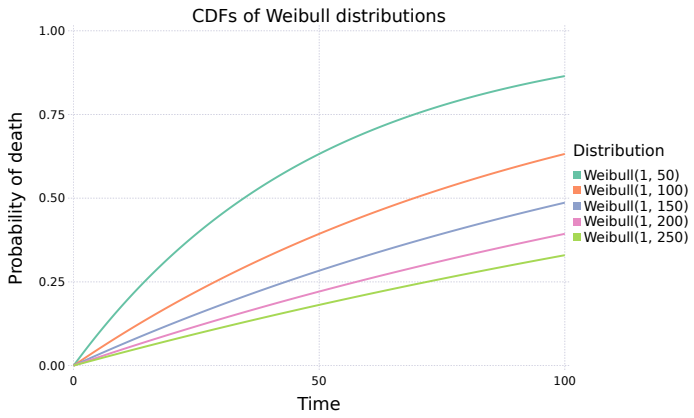

Figure 3: Examples of the functions that give the probability of death of a patient as a function of time after admission into the ICU.

We assume that, for all patients who are brought in, probability of death can be roughly modeled by the cumulative density function (CDF) of some Weibull distribution. Such distributions are characterized by two parameters, a shape parameter k and a scale parameter l, and the associated CDF is given by

$$F(x;k,l) \quad = \quad \begin{cases} 1-\exp\{-(x/l)^k\} & \text{for } x \geqslant 0, \\ 0 & \text{otherwise.} \end{cases}$$

More specifically, we assume that the probability of death for a given patient can be modeled in terms of an instance of this schema, where for each patient k and l are chosen randomly, with $k \sim \mathcal{U}(0.5,5)$ and $l \sim \mathcal{U}(50,250)$. See Figure 3 for some examples of such CDFs.

Also associated with each patient are two parameters, $a > 0$ and $b > 0$, that indicate the effect on probability of death of the right intervention and the effect on that probability of a wrong intervention, respectively. Specifically, where p_t is the probability of death of a given patient at time t, performing the right intervention at t lowers the probability of death to p_t/a while performing a wrong intervention at t raises that probability to $(1+p_t)/b$. Figure 4 illustrates these effects for the case where $a = b = 2$, $k = 1$, and $l = 50$.

We further assume that "what is wrong" with the patient can be expressed in terms of one parameter, α; we imagine that, at the time the patient is rushed into the ICU, her relevant medical status is known up to the value of this parameter. At that time, the only thing known about α is that it can take a value in $\{0,.1,.2,\ldots,1\}$, with none of these values initially being more likely than any other. To estimate the value of α, the doctor has to go on the test results that she receives, with one new result coming in per unit of time. The results are either "positive" or "negative," and the tests are probabilistically independent of each other and all have the same (unknown) probability of being positive. The hypothesis that $\alpha = x$ is to be interpreted as implying that the probability for any given test turning up positive equals x.

Pauker and Kassirer (1980) and Djulbegovic et al. (2014) found that the so-called threshold model, according to which physicians should decide to administer treatment if, and only if, the probability of disease is above a specified threshold, accurately predicts decision making in clinical practice (see also Djulbegovic & Elqayam 2017). In line with this finding, we will assume that the doctor must be at least 90 percent certain about a hypothesis before she is prepared to intervene.[11] It is further stipulated that only if the doctor comes to believe the

[11] Djulbegovic and colleagues found that the model which relates the threshold to the expected utility of treatment does best. (Actually, the best model was the one that took into account both expected utility and regret. However,

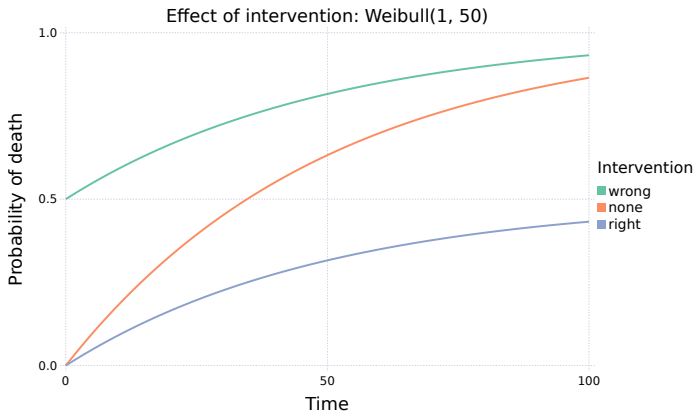

Figure 4: Example of the effect of right and wrong interventions: the orange graph is the probability of death of the patient over time if no intervention is performed; the green graph gives, for every point in time, the probability of death of the patient if at that point in time a wrong intervention is performed; the blue graph does the same for the correct intervention.

true hypothesis to a degree above that threshold will she perform the correct intervention; coming to believe a false hypothesis to a degree above the threshold will lead to an incorrect intervention, with any incorrect intervention having the same detrimental effect (as specified by the parameter b) on the probability that the patient will survive.

By now, we have described all structural properties of the context the doctor is to operate in as well as the various strategies (update rules) available to her in this context. Assuming that the doctor's goal is to save her patients' lives, which update rule should she use to update her degrees of belief for the various hypotheses concerning α (that $\alpha = 0$, that $\alpha = .1$, and so on)?

We saw earlier that, in a structurally identical probabilistic model, updating via one instance of EXPL tended to yield high-probability assignments to the truth faster than did Bayes' rule. On the other hand, we also noted that we should expect that instance of EXPL to have a higher error rate, in that it has a greater tendency to assign high probability to a false hypothesis than Bayes' rule does; the argument given for that did not depend on the exact value of c and so generalizes to all instances of EXPL. For reasons already mentioned, speed of convergence and accuracy are both important in the doctor's context: with every unit of time that passes, the probability that the patient will die goes up; but if the doctor acts upon a false hypothesis, her intervention will make the patient's prospects even worse. So the question just raised can be translated as: which update rules offers the best trade-off between speed and accuracy in the context at hand?

it is possible to model regret in terms of expected utility as well. Thanks to Shira Elqayam for useful discussion.) If one likes, one may assume that, for instance in view of possible long-term adverse effects of the intervention, in the context here at issue the expected utility of intervening exceeds that of non-intervention precisely if the doctor becomes at least 90 percent confident in one of the hypotheses about the value of α.

4.2 Method

We can think of this question as a constrained optimization problem.[12] Even though all relevant features of the setting have been characterized mathematically, this problem appears too complex for analytical methods to give much guidance. However, we can resort to an optimization technique known as "agent-based optimization," which is a form of genetic programming (Holland 1975; Koza 1992). This technique takes its cue from biological evolution, letting agents represent different solutions to a given problem, determining their "fitness" level (according to some criterion of fitness deriving from whatever problem needs to be solved), and then selecting the fittest agents, which are retained and / or allowed to reproduce in some pre-determined way and thereby provide the input population for the next round, in which the competition for survival or reproduction starts again. This can be repeated over and over, possibly as often as is needed to obtain a fixed point at which all agents represent the same solution (see, e.g., Barbati, Bruno, & Genovese 2012; also various papers in Czarnowski, Jędrzejowicz, & Kacprzyk 2013). As will be seen further on, using agent-based optimization has the additional advantage of allowing us to shed some light on how evolution may have helped to shape our inferential practices, and in particular may have led to the adoption of certain forms of explanatory reasoning.

The agent-based optimization procedure we conducted started with a population consisting of 200 "doctors," with 50 of them using Bayes' rule, 50 using an instance of EXPL, 50 using an instance of Good's rule, and 50 using an instance of Popper's rule; for the non-Bayesian updaters, the value of the explanation bonus c was chosen randomly per doctor, with $c \sim \mathcal{U}(0, 0.25)$. Then each doctor was assessed on the basis of treating 100 patients, where the relevant characteristics of the patients (survival probability, effects of right and wrong interventions, value of α) were chosen randomly and separately per patient, in the way specified above (so the survival probability was based on a randomly chosen Weibull distribution, with the parameters falling within the indicated bounds, and so on).

For each patient, the doctor had 100 units of time available. Per unit of time, the doctor received the outcome of a test, which was positive with a probability determined by the value of α as randomly chosen for the agent. At time 0, the doctor deemed all 11 value hypotheses for α equally likely, and probabilities were updated at each following time step on the basis of the test result received at that step, and using the update rule associated with the doctor. The doctor intervened as soon as the probability for one hypothesis exceeded the threshold of .9. If that probability was assigned to the *true* hypothesis, the doctor received the score of 1 minus the probability of death associated with the *right* intervention at the time the probability crossed the threshold; if that probability was assigned to a *false* hypothesis, the doctor received the score of 1 minus the probability of death associated with the *wrong* intervention at the time the probability crossed the threshold; and if *no* hypothesis was assigned a probability above the threshold during the 100 time steps, the doctor received the score of 1 minus the probability of death at the 100^{th} time step.

After each doctor had treated her 100 patients, her total score was calculated, and the 50 percent fittest doctors (the doctors with the highest average survival rate among their patients) were determined. These 100 doctors were retained for the next generation, and they were also allowed to replicate by having a copy of themselves in that generation. In every simulation, the previous steps were repeated for 100 generations, and in total 100 simulations were run. Figure 5 is a schematic representation of the steps involved in going from one generation to the next.

[12] *Constrained* optimization, because we are not looking for the optimal update rule per se (optimal given the context), but the one among the rules we are *considering* that best fits the context. This makes the current approach different from Anderson's (1990) rational analysis, which starts with what are basically the three steps from Arkes and colleagues summarized at the end of Section 3.3 but then adds as a further step a requirement to look for the optimal procedure *tout court* for achieving the individual's or group's goal.

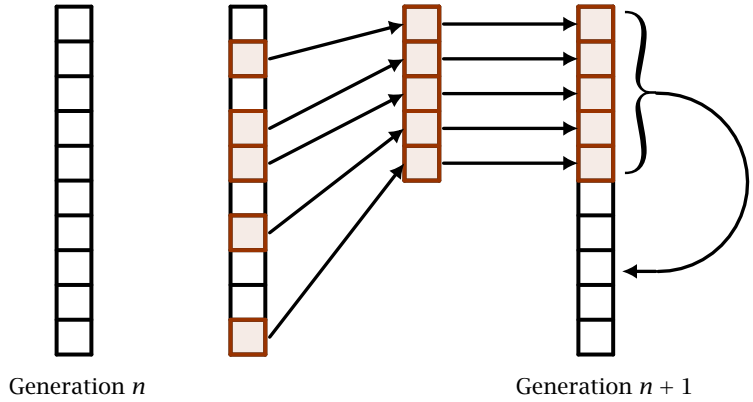

Figure 5: Schematic representation of the procedure used to go from generation n to generation $n+1$. First, the agents' fitness is determined; then the 50 percent fittest agents (shaded) are selected and retained for the next generation; finally, these agents are allowed to replicate, which yields the new generation.

This is an informal description of the simulations we conducted. Mathematical and computational details are relegated to the Jupyter notebook `Section4A.ipynb` that is part of the Supplementary Materials. Moreover, the program for the simulations was written in Julia, a new dynamic language for scientific computing (Bezanson et al. 2017), which reads almost like pseudo-code. So, going through the code in the notebook should further enhance understanding of the simulations.

4.3 Results

To give a first impression of the kind of results we obtained, Figure 6 shows for four randomly chosen simulations how the population evolved through the 100 generations. One can think of each of the four plots as a series of 100 stacked histograms, where these histograms show how many tokens of each agent type were present in the corresponding generation.

There are a number of observations we can immediately make. First, Bayesians (here in the sense of doctors using Bayes' rule) disappear fairly quickly. Second, while explanatory reasoning prevails in all four examples in some form or other, there is no one explanation-based update rule that clearly trumps the others. Third, EXPL users appear to be critically endangered during the initial stages of the shown simulations. In the examples in which they perish, they do so even faster than the Bayesians, but where they make it through the initial stage, they become the ultimate winners, pushing out of competition all other types of agents.

As for the overall results, we first consider the summary information about the last generations of all simulations, given in Figure 7. The left panel gives the counts of the type memberships among all $100 \times 200 = 20{,}000$ agents in the last generations. An observation from the above examples that clearly generalizes is that Bayesians are consistently out-competed into extinction. We also see that, although among the explanation-based rules there is none that always comes to dominate the field, there is a clear sense in which EXPL is the winner, with Popper's rule being a distant second; on average, users of Good's rule are sparsely represented in the last generation.

The right panel of Figure 7 plots the densities of the bonus values in the last generations, for the three agent types that attribute such values to hypotheses. The results are rather different for the different types. It is seen that the bonus values of the EXPL users are narrowly

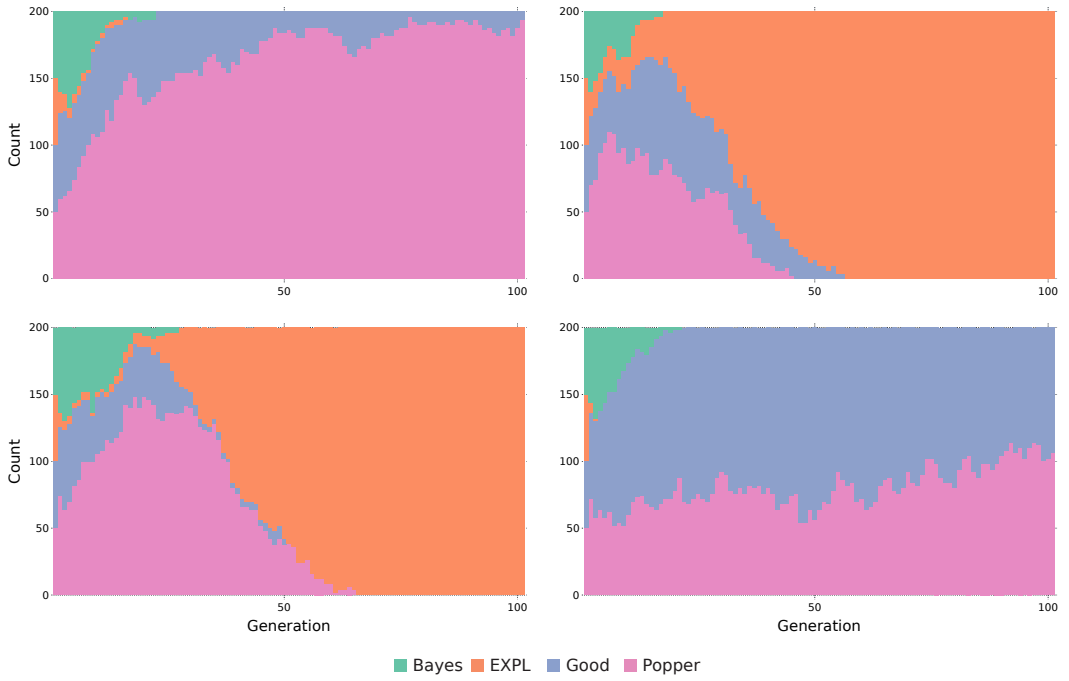

Figure 6: For four randomly chosen simulations, counts of agent types per generation. (See the text for further explanation.)

peaked at a value close to the bottom of the scale, while those of the Popperians are peaked—though not as narrowly—toward the opposite end of the scale. We used the `emmeans` package (Lenth 2018) for the statistical computing language R (R Core Team 2017) to obtain estimated marginals means (EMMs) for the bonus values, finding that the EMM for last-generation EXPL users was 0.010, 95 % CI $[0.010, 0.011]$; for last-generation users of Good's rule, the EMM was 0.117, 95 % CI $[0.116, 0.118]$; and for last-generation users of Popper's rule, the EMM was 0.227, 95 % CI $[0.226, 0.227]$.

To make our answer to the optimization question complete, we must compare the fitness levels—the average scores on the fitness function, which measures the likelihood of patient survival—of the different types of agents. A one-way ANOVA revealed that agent type had a

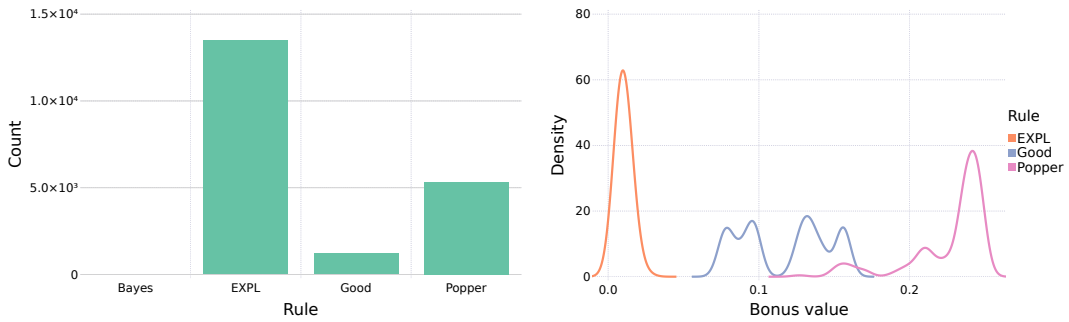

Figure 7: Count of agents per type in the last generations of the simulations (left), and densities of the bonus values of those agents, also per type (right).

significant effect on fitness level, $F(2, 19997) = 18.22$, $MSE = 0.0003$, $p < .0001$, $\eta^2 = .002$. Planned comparisons with Bonferroni correction showed that the mean fitness level of last-generation EXPL users (EMM = 0.873, 95% CI [0.873, 0.873]) differed significantly from the mean fitness levels of last-generation users of Good's rule (EMM = 0.871, 95% [0.870, 0.872]) as well as last-generation users of Popper's rule (EMM = 0.872, 95% CI [0.871, 0.872]), both ps $< .0001$. The mean fitness levels of the last two groups did not differ significantly.

While the effect of agent type on fitness level is small, the foregoing results can still be interpreted as showing that, on average, of every 10,000 patients admitted to the ICU, an EXPL doctor (using a bonus value for EXPL that is optimal for the current setting) would save 24 extra lives as compared to a doctor using Good's rule (again with the bonus value optimal for this rule in the current setting) and 14 extra lives as compared to a doctor using Popper's rule (again with optimal bonus value). Numerically, these may not be large differences. Yet it is hard to imagine a doctor who would *not* prefer EXPL (with suitable bonus value, and in the context at issue) over either of the other rules.

So far, the evolutionary perspective taken was only a means for answering the question of which rule would be ecologically most rational in our specific ICU setting—the answer being, as we just saw, that a doctor would be best off, on average, if she updated by EXPL, using a bonus value of around 0.01, and that therefore it is rational to use EXPL in this context. However, the same perspective also allows us to broaden the notion of environment to one in which agents are under evolutionary pressure and to consider what competitive advantages or disadvantages the various update rules may bring. Thus, instead of considering a single doctor at an ICU in the kind of setting specified above, we can study populations of such doctors who are in competition with each other—say, competition for an extension of their temporary contract, which has as a fringe benefit that it permits recruiting a friendly colleague who adheres to exactly the same updating mechanism—and where the selection is on the basis of how well they perform in comparison to others in the population. If we look at the full simulations again, and not only at the last generation in each, we see how populations of this kind evolve if at starting time there are four equally sized groups of agents, differentiated only by the update rules their members use.

Figure 8 gives relevant summary information. The left panel plots the percentage of agents of each type for every generation, averaged over all simulations. It shows that some of the observations made about the examples in Figure 6 hold generally. We see the strong tendency for EXPL users to drop dramatically in number shortly after the start. We also see that their fate remains in serious jeopardy for some time after the start—if they do not perish completely, as the examples showed may happen. But when they manage to survive that critical phase, they tend to quickly get the upper hand.

The right panel of Figure 8 plots the mean bonus values in each generation, for each of the three groups assigning such values. Because for the agents in all three groups the initial bonus values c were randomly drawn, with $c \sim \mathcal{U}(0, 0.25)$, all simulations start with an average bonus value of around 0.125 for each group. We see that the evolutionary process almost immediately drives down the average bonus value associated with the EXPL group to a level at which we found last-generation EXPL agents to assign bonuses. The other groups of explanation-based updaters appear to get by with average values that are closer to 0.125. So, the question at the beginning of each simulation appears to be whether for the EXPL group the average bonus value will go down *fast enough*, lest all EXPL users end up in the bottom half of least fit agents, which would eliminate them in the simulation.

It heavily weighs on the initial fitness of EXPL agents that they start out, on average, assigning a bonus that is so far from the apparent optimum. This can be seen by comparing the right panel of Figure 8 with Figure 9, which shows how the average fitness levels of the four groups develop over the course of the generations. One observes that whereas the average fitness levels of Good's rule and Popper's rule users are, over the whole course, quite close,

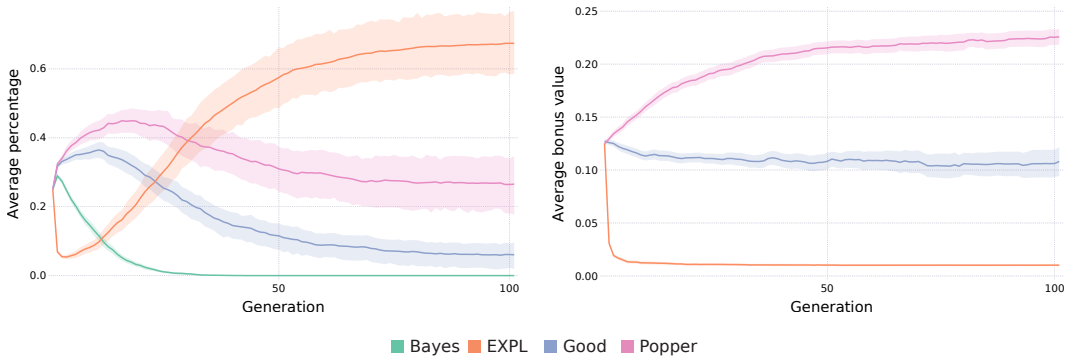

Figure 8: Percentage of agents per type of agents and per generation, averaged over all simulations (left), and mean bonus values per type of agents and per generation, also averaged over all simulations (right), both shown with 95 percent confidence bands.

with the average fitness level of Bayesians (for as long as they last) being almost consistently quite a bit below those of the aforementioned groups, the average fitness level of EXPL users is initially *far* below the levels of the other groups. But as a group, they tend to catch up quickly, and then are, on average, consistently fitter than the agents in the other groups.

It is further to be noticed that the difference in average fitness between EXPL users and the agents in either of the other groups of explanation-based updaters is small for the whole segment after the initial stage. But as has frequently been remarked, from an evolutionary viewpoint, small differences may be as advantageous as larger ones: often enough, a prey animal can escape a predator by being just a little faster than some other individuals in its group.

Perhaps the most important advantage of the explanation-based update rules these comparisons point at is that they offer an opportunity for adaptive learning—by increasing or decreasing the bonus for explanatory goodness—that Bayes' rule does not provide. How

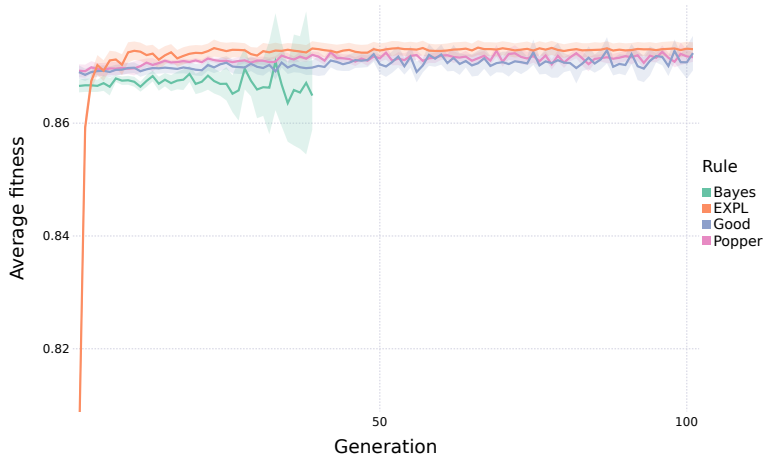

Figure 9: Average fitness per agent type (i.e., average probability that an ICU patient treated by a doctor of the given type will survive), over the generations, shown with 95 percent confidence bands.

much of a benefit this is is especially clear if we contrast Bayesians with EXPL users. At first, the latter lag well behind in terms of fitness. But EXPL users can adapt to the needs of the context (again, as a group, not at the individual level, at which bonus values are fixed). In our simulations, they were mostly able to do this fast enough to avoid extinction, and even to become fitter than their competitors. This is important, inasmuch as long-term survival is typically not just a matter of being well adapted to the environment one operates in but also of being able to adapt quickly to a new environment.

The point of the foregoing is not that one should always prefer one or the other instance of EXPL as an update rule. In fact, the Jupyter notebook `Section4B.ipynb` in the Supplementary Materials can be used to rerun the simulations reported in the above, but now using Gamma distributions instead of Weibull distributions to model probabilities of death. As can be seen in the notebook, in those simulations Popper's rule turns out best, with EXPL coming in second, Good's rule a distant third, and Bayesians (again) never making it to the last generation. Interested readers will have no difficulty tweaking the computer code to explore further variations still, but already the aforementioned result buttresses Elqayam's earlier-quoted remark that it would be a mistake to expect a one-size-fits-all norm of rationality.

More generally, it is important to be clear about the logic of the argument. According to its advocates, we are to follow Bayes' rule in each and every context. Not doing so would make us irrational, as the dynamic Dutch book and inaccuracy-minimization arguments are supposed to show. To refute this claim, it is enough to specify some context in which we are better off by following a non-Bayesian update rule. That leaves open the possibility that there are contexts in which Bayes' rule helps you achieve whatever it is you want to achieve more quickly or reliably or efficiently than would any explanation-based update rule or indeed any other rule—which is fine, given that our aim was to show that in *some* contexts using a form of explanatory reasoning is ecologically more rational than using Bayes' rule, and hence that Bayes' rule is not defensible as a *universally* valid principle of rationality; our aim was not to show that Bayes' rule is defensible under *no* circumstances, and that we should *always* adhere to some explanation-based update rule.[13]

5 Concluding remarks

We started by reviewing evidence indicating that the way people change their degrees of belief is influenced by explanatory factors, and that this may cause them to violate Bayesian norms of reasoning. We then asked *why* they would do that. The standard arguments in support of Bayesianism suggested an answer in terms of a bias, of something regrettable, even if perhaps not completely avoidable.

We showed those arguments to be fallacious, however. The key observation, generally put, was that even granting that failure to minimize next-step inaccuracy and / or Dutch-bookability are *really really bad*, nothing follows so long as it has not been ruled out that there is also a plus side to be considered, and that there is something *really really good* to be had on that side. And we did argue that people's tendency to rely on explanatory considerations in their

[13] Note that therefore it is also not a flaw of our simulations that they relied on assumptions (about the characteristics of the patients, the time available for intervention, the threshold of certainty for action, and so on) that were to some extent arbitrary. That would only be a problem if our aim had been to argue for a replacement of Bayes' rule by a different supposedly universally valid update rule. Meanwhile I should note that, although I had expected to find contexts in which Bayesians prevailed, in the many variations of the simulations that I tried out, that never happened. This is unfortunate to the extent that it could have given valuable clues as to how to distinguish formally between the kinds of context in which some explanation-based rule may be expected to do best and those in which Bayes' rule may be expected to do best. (Naturally, when there is no sensible way in which the hypotheses under consideration could be said to explain the data one is updating on, then all the explanation-based update rules boil down to Bayes' rule and so that becomes automatically one's best choice, at least supposing the choice to be limited to the rules considered in this paper.)

belief updating can have clear advantages. These advantages can be cognitive—helping us find out the truth faster—as well as practical, such us helping us save lives. We are truth seekers, but our time is limited, and sacrificing some accuracy in return for a generally faster convergence to the truth can be an acceptable compromise.

Whether it is, and to which extent we should be willing to compromise, depends on the environment we are in as well as on our goal in that environment. Therefore, the claim was that explanatory reasoning is rational in some environments—rational in a sense that lets us focus on how well a cognitive strategy (e.g., an update rule) is adapted to the local environment in which it is deployed rather than on how well the strategy complies to certain internal standards, such as consistency or probabilistic coherence.

The problem for the traditional Bayesian approach to updating, and to rationality in general, that we have aimed to highlight is not that it posits an ideal that we would not even know how to approximate. Instead, the problem is that, at least in some situations, the use of Bayes' rule is simply *not* ideal. In Section 4, we saw a straightforward example of this: use of Bayes' rule led to the survival of fewer patients, on average, than the use of some version of explanatory reasoning. We also saw there that what helped some of those versions do so well in that context is that they are highly adaptable by having an adjustable parameter, which allows for contextual fine-tuning. Bayes' rule lacks this functionality. More generally, given that we inhabit an ever changing world and can within a relatively short time span find ourselves in contexts that pose very different challenges, reasoning should not just be well-adapted to the context we are in at a given point in time; it should also be easily *adaptable* to whichever context we may, just moments later, find ourselves in (Schurz & Thorn 2016). Insisting on the unique rationality of Bayes' rule would make us needlessly inflexible. If we are instead permitted to take on board instances of EXPL, Good's rule, Popper's rule, and other rules perhaps (including Bayes' rule), that would greatly facilitate quick adaptation, by switching from EXPL to Good's rule, for instance, and / or by adjusting the explanation bonus.

It is fair to say that the sort of psychological approach to rationality that we have been assuming in this paper has been mostly concerned with arguing for the ecological validity of so-called fast and frugal heuristics, rather than that of the kind of higher-level cognitive principles that were compared in this paper. A key insight of the work of Gigerenzer and his various collaborators is that the use of simple heuristics will often lead to better outcomes than the use of more complicated principles of reasoning. However, nothing in Gigerenzer et al.'s work suggests that the notion of ecological rationality only applies to heuristics. Indeed, we saw in the previous section how clear sense can be given to the notion of one update rule matching the environment better than another update rule.[14]

Our main method was that of computer simulations, in particular, agent-based modeling. The framework that was used for the simulations is highly flexible and allows of almost endless variations. One could explore still further distributions for modeling probability of death, different ways of modeling the effects of an intervention, richer and more varied sources of information on which agents update, and also different modes of reproduction (Bäck 1996, Ch. 2). Studying such variations provides a path to probing the *robustness* of update rules, which indicates how broadly applicable they are (Gigerenzer 2001, p. 47). Part of the conception of rationality that we have been assuming is that rational agents are able to

[14] Admittedly, concerns of computational intractability that Gigerenzer and others have voiced over Bayesian principles would seem to apply equally to the explanation-based rules studied in this paper; in fact, given that they appear slightly more complex than Bayes' rule (in view of the additional computational work needed to assign bonuses), the latter principles may give even more cause for concern in this respect. But for all that has been said, the formal rules that we studied may in actuality be implemented by informal heuristics, or at least the former may be approximated by the latter. And to shed light on the informal heuristics by means of computer simulations, there may be no alternative to relying on such formal approximations. (See Schurz & Thorn 2016 and Schurz 2019 for various other examples of the same approach within the paradigm of ecological rationality.) However, I would like to leave processing issues related to explanatory reasoning as a topic for another paper.

pick the right tools for the right situation; just considering updating on incoming information, this may include knowing when best to use a particular instance of EXPL, say, and knowing when to use Bayes' rule, or some other rule still.[15] But this is not to say that broad applicability is not a boon for update rules. If an update rule matches a large range of contexts, or at least matches it better than all other known update rules, then that offers agents a good opportunity to become proficient users of that rule and lets them benefit in new contexts from the experience they have gained with the rule in earlier contexts.

Another variation one could consider is to build in a mechanism of social learning, for instance, in the manner of the well-known Hegselmann–Krause model (Hegselmann & Krause 2002, 2006, 2009; see also Chen, Glass, & McCartney 2018). Instead of only updating on evidence, agents can take into account the degrees of belief of others in their community, perhaps of those they regard as their peers, or as experts on the matter at issue. The importance of social learning in relation to the ecological conception of rationality has been repeatedly emphasized by Gigerenzer (e.g., in his 2000). Social learning of Bayesians and of EXPL updaters has been explored by Douven and Wenmackers (2017), though only in the abstract, looking at speed of convergence and accuracy, but without any contextualization, and so without being able to determine what would count as a reasonable speed–accuracy trade-off.

Furthermore, one could extend the competition to other update rules. For instance, one could bring in the type of meta-inductive reasoners canvassed by Schurz (2008b, 2009, 2019; also Schurz & Thorn 2016), which keep track of the success rates of other reasoners and use that information to inform their own updating. Schurz (2019, Ch. 6) introduces a variety of meta-inductive strategies and discusses what difference they may make to predictive performance. It would be straightforward to implement and compare those strategies in the context of our evolutionary computations.

The explanation-based update rules were built on certain measures of explanatory goodness, but there are other such measures (e.g., Hintikka 1968; Schupbach & Sprenger 2011) as well as closely related measures, notably measures of coherence (Bovens & Hartmann 2002; Olsson 2002; Douven & Meijs 2007; Glass 2007, 2012, 2018) and measures of the degree of competition among hypotheses (Schupbach & Glass 2017; Glass 2019), and considering rules using such alternatives is another avenue for future research.[16] Indeed, contributions to the debate about explanatory reasoning tend to be noncommittal on theories of explanation, and in general there has been little contact between the debate about rational updating and the debate about explanation (Kuipers 2000 and Lipton 2004 are exceptions). But it would be worth investigating whether the various proposals concerning explanation that have been made in the philosophy of science can be made formal enough to compare them in the kind of computational setting developed in the above. For instance, might a suitably formalized causal model of explanation (perhaps building on Pearl's 2000 seminal work) do better in the kind of simulations we ran than an equally suitably formalized unificationist model (perhaps building on the aforementioned measures of coherence)? It would be an interesting finding if, say, inferences to the best causal explanation turned out to be generally more successful, or in some contexts more successful, than inferences to the best unifying account.[17]

Last but not least, empirical research on explanatory reasoning served as a source of inspiration for the simulations we conducted in this paper. Registered systematic violations of Bayesian updating led us to consider explanation-based update rules, and experimental results reported in Douven and Schupbach (2015a, 2015b) led to the proposal of Good's rule

[15] There thus appears to be a connection between rational updating and meta-cognition (Thompson, Prowse Turner, & Pennycook 2011; Ackerman & Thompson 2017, 2018). I only flag this here to set it aside for future work.

[16] One can of course have doubts about the measures of explanatory goodness that were used in this paper (see, e.g., Glymour 2015). However, I am not aware of arguments to the effect that our explanatory judgments cannot possibly be modeled by a probabilistic measure.

[17] I owe this suggestion to an anonymous referee.

and Popper's rule. We found normative reasons for preferring these rules over Bayes' rule, at least in certain contexts. It would now be worth conducting further experimental work to investigate whether any of the non-Bayesian rules we looked at is able to more generally predict people's deviations from Bayesian reasoning. And with some luck that empirical work could in turn help us to formulate versions of explanation-based update rules that are still formal but at the same time more realistic than EXPL, Good's rule, and Popper's rule.[18]

Bayesianism enjoys widespread popularity among theorists concerned with rationality. Our results were not meant to show that this popularity is undeserved; adopting Bayesian norms may be the right thing to do in certain contexts. But Bayesian *imperialism*, according to which those norms should govern our behavior, cognitive and practical, in each and every context, is misguided. Most critically, this imperialism has gone at the expense of proposals that assign a guiding role to explanatory considerations. It was shown how taking into account such considerations can carry important benefits, which should be reason to give the said proposals a fairer hearing than philosophers and also many psychologists have been willing to do so far.

Acknowledgments

I am greatly indebted to Shira Elqayam, Patricia Mirabile, Christopher von Bülow, and two anonymous referees for valuable comments on previous versions. Thanks also to audiences at the Frankfurt School of Finance and Management, the IHPST (Paris), and the VU University (Amsterdam) for useful questions and remarks.

References

Ackerman, R., & Thompson, V. A. (2017). Meta-reasoning: Monitoring and control of thinking and reasoning. *Trends in Cognitive Sciences*, *21*, 607–617.

Ackerman, R., & Thompson, V. A. (2018). Meta-reasoning: Shedding meta-cognitive light on reasoning research. In L. J. Ball & V. A. Thompson (eds.), *International handbook of thinking and reasoning* (pp. 1–15). London: Routledge.

Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale NJ: Erlbaum.

Arkes, H. R., Gigerenzer, G., & Hertwig, R. (2016). How bad is incoherence? *Decision*, *3*, 20–39.

Bäck, T. (1996). *Evolutionary algorithms in theory and practice*. Oxford: Oxford University Press.

Barbati, M., Bruno, G., & Genovese, A. (2012). Applications of agent-based models for optimization problems: A literature review. *Expert Systems with Applications*, *39*, 6020–6028.

Berg, N., Biele, G., & Gigerenzer, G. (2016). Consistent Bayesians are no more accurate than non-Bayesians: Economists surveyed about PSA. *Review of Behavioral Economics*, *3*, 189–219.

Bes, B., Sloman, S., Lucas, C. G., & Raufaste, E. (2012). Non-Bayesian inference: Causal structure trumps correlation. *Cognitive Science*, *36*, 1178–1203.

Bezanson, J., Edelman, A., Karpinski, S., & Shah, V. B. (2017). Julia: A fresh approach to numerical computing. *SIAM Review*, *59*, 65–98.

Bickel, J. E. (2010). Scoring rules and decision analysis education. *Decision Analysis*, *7*, 346–357.

Bovens, L., & Hartmann, S. (2003). *Bayesian epistemology*. Oxford: Oxford University Press.

[18] Thanks to an anonymous referee for suggesting this.

Boyd, R. (1990). Realism, approximate truth, and philosophical method. In C. Wade Savage (ed.), *Scientific theories* (pp. 355–391). Minneapolis MN: University of Minnesota Press.
Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, *78*, 1–3.
Bunt, H. C., & Black, W. J. (2000). *Abduction, belief, and context in dialogue: Studies in computational pragmatics*. Cambridge MA: MIT Press.
Chen, S., Glass, D. H., & McCartney, M. (2018). Two-dimensional opinion dynamics in social networks with conflicting beliefs. *AI and Society*, in press.
Cooke, R. M. (1991) *Experts in uncertainty*. Oxford: Oxford University Press.
Costello, F., & Watts, P. (2016). People's conditional probability judgments follow probability theory (plus noise). *Cognitive Psychology*, *89*, 106–133.
Costello, F., & Watts, P. (2018). Invariants in probabilistic reasoning. *Cognitive Psychology*, *100*, 1–16.
Czarnowski, I., Jędrzejowicz, P., & Kacprzyk J. (2013). *Agent-based optimization*. Berlin: Springer.
Djulbegovic, B., & Elqayam, S. (2017). Many faces of rationality: Implications of the great rationality debate for clinical decision-making. *Journal of Evaluation in Clinical Practice*, *23*, 915–922.
Djulbegovic, B., Elqayam, S., Reljic, T., Hozo, I., Miladinovic, B., Tsalatsanis, A., Kumar, A., Beckstead, J., Taylor, S., & Cannon-Bowers, J. (2014). How do physicians decide to treat: An empirical evaluation of the threshold model. *BMC Medical Informatics and Decision Making*, *14*, 47, http://www.biomedcentral.com/1472-6947/14/47.
Douven, I. (1999). Inference to the best explanation made coherent. *Philosophy of Science*, *66*, S424–S435.
Douven, I. (2002). Testing inference to the best explanation. *Synthese*, *130*, 355–377.
Douven, I. (2013). Inference to the best explanation, Dutch books, and inaccuracy minimisation. *Philosophical Quarterly*, *69*, 428–444.
Douven, I. (2016a). *The epistemology of indicative conditionals*. Cambridge: Cambridge University Press.
Douven, I. (2016b). Explanation, updating, and accuracy. *Journal of Cognitive Psychology*, *28*, 1004–1012.
Douven, I. (2017). Inference to the best explanation: What is it? And why should we care? In T. Poston & K. McCain (eds.), *Best explanations: New essays on inference to the best explanation* (pp. 4–22). Oxford: Oxford University Press.
Douven, I. (2019a). Can the evidence for explanatory reasoning be explained away? *IEEE Transactions on Cognitive and Developmental Systems*, in press.
Douven, I. (2019b). Optimizing group learning: An evolutionary computing approach. *Artificial Intelligence*, *275*, 235–251.
Douven, I. (2019c). Scoring in context. *Synthese*, in press.
Douven, I., Elqayam, S., Singmann, H., & van Wijnbergen-Huitink, J. (2018). Conditionals and inferential connections: A hypothetical inferential theory. *Cognitive Psychology*, *101*, 50–81.
Douven, I., Elqayam, S., Singmann, H., & van Wijnbergen-Huitink, J. (2019). Conditionals and inferential connections: Toward a new semantics. *Thinking and Reasoning*, in press.
Douven, I., & Meijs, W. (2007). Measuring coherence. *Synthese*, *156*, 405–425.
Douven, I., & Mirabile, P. (2018). Best, second-best, and good-enough explanations: How they matter to reasoning. *Journal of Experimental Psychology: Language, Memory, and Cognition*, *28*, 1099–1124.
Douven, I., & Schupbach, J. N. (2015a). The role of explanatory considerations in updating. *Cognition*, *142*, 299–311.

Douven, I., & Schupbach, J. N. (2015b). Probabilistic alternatives to Bayesianism: The case of explanationism. *Frontiers in Psychology*, *6*, doi: 10.3389/fpsyg.2015.00459.

Douven, I., & Wenmackers, S. (2017). Inference to the best explanation versus Bayes' rule in a social setting. *British Journal for the Philosophy of Science*, *68*, 535–570.

Earman, J. (1992). *Bayes or bust?* Cambridge MA: MIT Press.

Eberhardt, F., & Danks, D. (2011). Confirmation in the cognitive sciences: The problematic case of Bayesian models. *Minds and Machines*, *21*, 389–410.

Eddy, D. M. (1982). Probabilistic reasoning in clinical medicine: Problems and opportunities. In D. Kahneman, P. Slovic, & A. Tversky (eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 249–267). Cambridge: Cambridge University Press.

Edwards, W. (1968). Conservatism in human information processing. In B. Kleinmuntz (ed.), *Formal representation of human judgment* (pp. 17–52). New York: Wiley.

Elqayam, S. (2011). Grounded rationality: A relativist framework for normative rationality. In K. I. Manktelow, D. E. Over, & S. Elqayam (eds.), *The science of reason* (pp. 397–420). Hove UK: Psychology Press.

Elqayam, S. (2012). Grounded rationality: Descriptivism in epistemic context. *Synthese*, *189*, 39–49.

Elqayam, S. (2018). New psychology of reasoning. In L. J. Ball & V. A. Thompson (eds.), *International handbook of thinking and reasoning* (pp. 130–150). London: Routledge.

Elqayam, S., & Evans, J. St. B. T. (2011). Subtracting 'ought' from 'is': Descriptivism versus normativism in the study of human thinking. *Behavioral and Brain Sciences*, *34*, 233–248.

Elqayam, S., & Evans, J. St. B. T. (2013). Rationality in the new paradigm: Strict versus soft Bayesian approaches. *Thinking & Reasoning*, *19*, 453–470.

Evans, J. St. B. T., & Over, D. E. (1996). *Rationality and reasoning*. Hove UK: Psychology Press.

Fantl, J., & McGrath, M. (2009). *Knowledge in an uncertain world*. Oxford: Oxford University Press.

Fischhoff, B., & Lichtenstein, S. (1978). Don't attribute this to Reverend Bayes. *Psychological Bulletin*, *85*, 239–243.

Gigerenzer, G. (2000). *Adaptive thinking: Rationality in the real world*. New York: Oxford University Press.

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review*, *103*, 650–669.

Gigerenzer, G., Hertwig, R., & Pachur T. (2011). *Heuristics: The foundations of adaptive behavior*. New York: Oxford University Press.

Gigerenzer, G., Todd, P. M, & the ABC Research Group (1999). *Simple heuristics that make us smart*. Oxford: Oxford University Press.

Glass, D. H. (2007). Coherence measures and inference to the best explanation. *Synthese*, *157*, 275–296.

Glass, D. H. (2012). Inference to the best explanation: Does it track truth? *Synthese*, *185*, 411–427.

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

Glass, D. H. (2019). Competing hypotheses and abductive inference. *Annals of Mathematics and Artificial Intelligence*, in press.

Glymour, C. (2015). Probability and the explanatory virtues. *British Journal for the Philosophy of Science*, *66*, 591–604.

Goldstein, D. G., & Gigerenzer, G. (2002). Models of ecological rationality: The recognition heuristic. *Psychological Review*, *109*, 75–90.

Good, I. J. (1960). Weight of evidence, corroboration, explanatory power, information and the utility of experiment. *Journal of the Royal Statistical Society*, *B22*, 319–331.

Hammerton, M. (1973). A case of radical probability estimation. *Journal of Experimental Psychology*, *101*, 252–254.

Harman, G. H. (1965). The inference to the best explanation. *Philosophical Review*, *74*, 88–95.

Hastie, R., & Pennington, N. (2000). Explanation-based decision making. In T. Connolly, H. R. Arkes, & K. R. Hammond (eds.), *Judgment and decision making: An interdisciplinary reader* (pp. 212–228). Cambridge: Cambridge University Press.

Hegselmann, R., & Krause, U. (2002). Opinion dynamics and bounded confidence: Models, analysis, and simulations. *Journal of Artificial Societies and Social Simulation*, *5*, http://jasss.soc.surrey.ac.uk/5/3/2.html.

Hegselmann, R., & Krause, U. (2006). Truth and cognitive division of labor: First steps towards a computer aided social epistemology. *Journal of Artificial Societies and Social Simulation*, *9*, http://jasss.soc.surrey.ac.uk/9/3/10.html.

Hegselmann, R., & Krause, U. (2009). Deliberative exchange, truth, and cognitive division of labour: A low-resolution modeling approach. *Episteme*, *6*, 130–144.

Hintikka, J. (1968). The varieties of information and scientific explanation. In B. van Rootselaar & J. F. Staal (eds.), *Logic, methodology, and philosophy of science III* (pp. 151–171). Amsterdam: North-Holland.

Hobbs, J. R. (2004). Abduction in natural language understanding. In L. Horn & G. Ward (eds.), *Handbook of pragmatics* (pp. 724–741). Oxford: Blackwell.

Holland, J. (1975). *Adaptation in natural and artificial systems*. Cambridge MA: MIT Press.

Jeffrey, R. C. (1965). *The logic of decision*. Chicago: University of Chicago Press.

Johnson, S. G. B., Rajeev-Kumar, G., & Keil, F. (2016). Sense-making under ignorance. *Cognitive Psychology*, *89*, 39–70.

Johnston, A. M., Johnson, S. G. B., Koven, M. L., & Keil, F. C. (2017). Little Bayesians or little Einsteins? Probability and explanatory virtue in children's inferences. *Developmental Science*, *20*, e12483, https://doi.org/10.1111/desc.12483.

Joyce, J. (1998). A nonpragmatic vindication of probabilism. *Philosophy of Science*, *65*, 575–603.

Koehler, D. J. (1991). Explanation, imagination, and confidence in judgment. *Psychological Bulletin*, *110*, 499–519.

Koslowski, B. (2018). Abductive reasoning and explanation. In L. J. Ball & V. A. Thompson (eds.), *International handbook of thinking and reasoning* (pp. 366–382). London: Routledge.

Kowlowski, B., Marasia, J., Chelenza, M., & Dublin, R. (2008). Information becomes evidence when an explanation can incorporate it into a causal framework. *Causal Development*, *23*, 472–487.

Koza, J. (1992). *Genetic programming: On the programming of computers by means of natural selection*. Cambridge MA: MIT Press.

Kuipers, T. A. F. (2000). *From instrumentalism to constructive realism*. Dordrecht: Kluwer.

Ladyman, J., Douven, I., Horsten, L., & van Fraassen, B. C. (1997). A defence of van Fraassen's critique of abductive reasoning: Reply to Psillos. *Philosophical Quarterly*, *47*, 305–321.

Lenth, R. (2018). `emmeans`: *Estimated marginal means, aka least-squares means*. `R` package version 1.1.3, http://cran.r-project.org/package=emmeans.

Lewis, D. K. (1980). A subjectivist's guide to objective chance. In R. C. Jeffrey (ed.), *Studies in inductive logic and probability, vol. 2* (pp. 263–293). Berkeley CA: University of California Press.

Lewis, D. K. (1999). Why conditionalize? In his *Papers in metaphysics and epistemology* (pp. 403–407). Cambridge: Cambridge University Press.

Lipton, P. (2004). *Inference to the best explanation* (2nd ed.). London: Routledge.

McCain, K., & Poston, T. (2014). Why explanatoriness is evidentially relevant. *Thought*, *3*, 145–153.

Maher, P. (1993). *Betting on theories*. Cambridge: Cambridge University Press.
Marks, D. F., & Clarkson, J. K. (1972). An explanation of conservatism in the bookbag-and-pokerchips situation. *Acta Psychologica*, *36*, 145–160.
Oaksford, M., & Chater, N. (2007). *Bayesian rationality*. Oxford: Oxford University Press.
Oaksford, M., & Chater, N. (2013). Dynamic inference and everyday conditional reasoning in the new paradigm. *Thinking & Reasoning*, *19*, 346–379.
Olsson, E. J. (2002). What is the problem of coherence and truth? *Journal of Philosophy*, *99*, 246–272.
Over, D. E. (2009). New paradigm psychology of reasoning. *Thinking & Reasoning*, *15*, 431–438.
Pauker, S. G, & Kassirer, J. P. (1980). The threshold approach to clinical decision making. *New England Journal of Medicine*, *302*, 1109–1117.
Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge: Cambridge University Press.
Pennington, N., & Hastie, R. (1988). Explanation-based decision making: The effects of memory structure on judgment. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*, 521–533.
Pennington, N., & Hastie, R. (1992). Explaining the evidence: Tests of the story-model for juror decision making. *Journal of Personality and Social Psychology*, *62*, 189–206.
Pennington, N., & Hastie, R. (1993). Reasoning in explanation-based decision making. *Cognition*, *49*, 123–163.
Pettigrew, R. (2016). *Accuracy and the laws of credence*. Oxford: Oxford University Press.
Phillips, L. D., & Edwards, W. (1966). Conservatism in a simple probability inference task. *Journal of Experimental Psychology*, *72*, 346–354.
Popper, K. R. (1959). *The logic of scientific discovery*. London: Hutchinson.
Poston, T. (2014). *Reason and explanation*. New York: Palgrave Macmillan.
Psillos, S. (1999). *Scientific realism: How science tracks truth*. London: Routledge.
R Core Team (2017). R: *A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, http://www.R-project.org/.
Rinard, S. (2017). No exception for belief. *Philosophy and Phenomenological Research*, *94*, 121–143.
Rittle-Johnson, B., & Loehr, A. M. (2017). Eliciting explanations: Constraints on when self-explanation aids learning. *Psychonomic Bulletin & Review*, *24*, 1501–1510.
Rosenkrantz, R. D. (1992). The justification of induction. *Philosophy of Science*, *59*, 527–539.
Schum, D. A., & Martin, A. W. (1982). Formal and empirical research on cascaded inference in jurisprudence. *Law and Society Review*, *17*, 105–151.
Schupbach, J. N. (2017). Inference to the Best Explanation, Cleaned Up and Made Respectable. In T. Poston & K. McCain (eds.), *Best explanations: New essays on inference to the best explanation* (pp. 39–61). Oxford: Oxford University Press.
Schupbach, J. N., and Glass, D. H. (2017). Hypothesis competition beyond mutual exclusivity. *Philosophy of Science*, *84*, 810–824.
Schupbach, J. N., & Sprenger, J. (2011). The logic of explanatory power. *Philosophy of Science*, *78*, 105–127.
Schurz, G. (2008a). Patterns of abduction. *Synthese*, *164*, 201–234.
Schurz, G. (2008b). The meta-inductivist's winning strategy in the prediction game: A new approach to Hume's problem. *Philosophy of Science*, *75*, 278–305.
Schurz, G. (2009). Meta-induction and social epistemology. *Episteme*, *6*, 200–220.
Schurz, G. (2019). *Hume's problem solved: The optimality of meta-induction*. Cambridge MA: MIT Press.
Schurz, G., & Hertwig, R. (2018). Cognitive success: A consequentialist account of rationality and cognition. *Topics in Cognitive Science*, in press.

Schurz, G., & Thorn, P. D. (2016). The revenge of ecological rationality: Strategy-selection by meta-induction within changing environments. *Minds and Machines*, *26*, 31–59.
Sidney, P. G., Hattikudur, S., & Alibali, M. W. (2015). How do contrasting cases and self-explanation promote learning? Evidence from fraction division. *Learning and Instruction*, *40*, 29–38.
Simon, H. A. (1982). *Models of bounded rationality, vol. 1*. Cambridge MA: MIT Press.
Sloman, S. A. (1994). When explanations compete. *Cognition*, *52*, 1–21.
Sloman, S. A. (1997). Explanatory coherence and the induction of properties. *Thinking & Reasoning*, *3*, 81–110.
Teller, P. (1973). Conditionalization and observation. *Synthese*, *26*, 218–258.
Thompson, V. A., Prowse Turner, J. A., & Pennycook, G. (2011). Intuition, reason, and metacognition. *Cognitive Psychology*, *63*, 107–140.
Todd, P. M., & Brighton, H. (2016). Building the theory of ecological rationality. *Minds and Machines*, *26*, 9–30.
Todd, P. M., & Gigerenzer, G. (2012). *Ecological rationality: Intelligence in the world*. Oxford: Oxford University Press.
Tregear, M. (2004). Utilising explanatory factors in induction? *British Journal for the Philosophy of Science*, *55*, 505–519.
Trpin, B., & Pellert, M. (2019). Inference to the best explanation in uncertain evidential situations. *British Journal for the Philosophy of Science*, in press.
van Fraassen, B. C. (1989). *Laws and symmetry*. Oxford: Oxford University Press.
Weisberg, J. (2009). Locating IBE in the Bayesian framework. *Synthese*, *167*, 125–143.
Williams, J. J., & Lombrozo, T. (2010). The role of explanation in discovery and generalization: Evidence from category learning. *Cognitive Science*, *34*, 776–806.
Williams, J. R. G. (2012). Generalized probabilism: Dutch books and accuracy domination. *Journal of Philosophical Logic*, *41*, 811–840.
Williamson, T. (2018). *Doing philosophy*. Oxford: Oxford University Press.