Abduction: Theory and evidence

Igor Douven

Abduction: Theory and Evidence

Igor Douven
IHPST / Panthéon–Sorbonne University / CNRS
igor.douven@univ-paris1.fr

Abstract

This chapter looks at new theoretical work on abduction, with a special focus on arguments concerning the normative status of abduction, as well as at empirical results relevant to the question of whether theories of abduction are descriptively adequate.

Keywords: abduction; Bayes' rule; belief change; computer simulations; explanatory reasoning; inference; probability.

1 Introduction

Broadly understood, abduction is the idea that explanatory considerations have confirmation-theoretic significance. What this means, to a first approximation, is that whenever we wonder how much confidence to invest in a hypothesis or theory, given the available evidence, we should *also* consider the question of how well the hypothesis or theory *explains* that evidence. Suppose, for instance, that we conduct an experiment and the results allow us to eliminate a number of theories in the relevant domain but leave still more than one contender in the running. Then if one of the remaining theories is a clearly better explanation of our experimental results than the others are, that is reason to put more confidence in the former than in any of the other theories; according to some authors, it is even reason to infer that the former is correct.

Abduction is different from the more widely studied inference form of deduction, if only because an abductive inference is revisable: we may receive additional evidence, and then some other theory may best explain our evidence, in which case we may become more confident, or infer categorically, that this other theory is correct. Abduction is also different from induction, another form of inference that is revisable. The key difference is that induction, as commonly understood, exploits only frequency information whereas abduction relies crucially on judgments of explanation quality (which, note, is not to exclude that these judgments may rely, at least partly, on frequency information).

Until at least 1980, philosophers of science and epistemologists took abduction more or less for granted. This changed with the advent of Bayesianism, which came to dominate thinking about rationality in the 1980s and 1990s, even to the extent that any rule apart from Bayes' rule (see below for details) came to appear suspect. The most general point of critique Bayesians raised against abduction was that whereas their position builds on a precise mathematical machinery, abduction is no more than a slogan. And even if abduction can be made formally precise, then it could still only be subservient to Bayes' rule, for instance, by recruiting explanatory intuitions in order to help determining prior probabilities, or by functioning as a heuristic shortcut to approximate probabilities whose exact calculation would take more time and effort than should be spent given the use

case at issue. Any formally precise version of abduction that is more aspiring is doomed, according to Bayesians, if not because it makes the user "Dutch-bookable" (i.e., open to engaging in sets of bets that she is guaranteed to lose), then because it makes the user's degrees of belief (i.e., subjective probabilities) more inaccurate than they would be were the user a Bayesian.

In view of these criticisms, should we still care about abduction? We should, for at least two reasons. First, there is evidence that people do reason abductively, and that they do so in ways that lead them to violate Bayes' rule; that makes the study of abduction worthwhile at least from a psychologist's standpoint. Second, there is recent work casting doubt on the Bayesian arguments according to which abduction violates norms of good reasoning. Among other concerns raised about these arguments, friends of abduction have countered that even if abduction has the flaws Bayesians attribute to it, there is reason to suspect that abduction can offer benefits in return which may more than make up for those flaws.

Section 2 looks at various ways in which authors have proposed to make the broad idea of abduction precise. Section 3 presents empirical evidence bearing on abduction. Section 4 critically discusses the main arguments that have been leveled against abduction. And Section 5, finally, canvasses a recent defense of the idea that, in the right kind of circumstances, abduction is a rational mode of reasoning.

2 What is abduction?

Above, abduction was characterized *broadly*. In the literature, one finds various more precise statements of this mode of inference. According to what is probably the most common characterization, abduction licenses the acceptance of a hypothesis on the basis that it best explains the available evidence (e.g., Psillos, 2004, 83). As various authors have pointed out, however, this characterization is unsatisfactory for more than one reason. A first reason is that, thus conceived, abduction authorizes an absolute judgment—accepting a hypothesis as true—on the basis of a relative one, to wit, that the hypothesis better explains the evidence than the other *available* candidate explanations, which will typically not include all potential explanations of the evidence (van Fraassen, 1989, Ch. 6). A second reason why the previous characterization has been deemed inadequate is that in cases in which the best explanation of our evidence is still a poor one, or *is* satisfactory but hardly more so than the second-best explanation, an inference to that best explanation would pre-theoretically appear unwarranted.

These concerns have inspired authors to come up with more refined proposals. For instance, Kuipers (1992) has addressed the first concern by proposing a reformulation of abduction according to which it licenses the inference to the conclusion that the best explanation is *closer to the truth* than the other available candidate explanations. And in response to the second concern, Lipton (1993) strengthens the standard definition of abduction by adding to it the requirement that the best explanation be both sufficiently good and sufficiently much better than its closest rival.

A more general concern that has been raised about abduction is that it lacks precision, whether in its standard formulation or in the versions of Kuipers and Lipton, and that thereby it contrasts unfavorably with Bayes' rule, which is its main contender. Admittedly, Bayes' rule comes as a precise mathematical formula, in comparison with which abduction can easily appear as a vague suggestion. However, already van Fraassen (1989, Ch. 6) proposed a probabilistic version of abduction, and recently a number of variants of that version have appeared in the literature.

According to Bayes' rule, a rational person updates her personal (or subjective) probabilities

upon the receipt of new information E by setting, for all propositions H expressible in her language,

$$\Pr_E(H) \;=\; \Pr(H \mid E) \;=\; \frac{\Pr(H)\,\Pr(E \mid H)}{\Pr(E)},$$

where $\Pr(\cdot)$ is the person's probability function right before she receives E and $\Pr_E(\cdot)$ her probability function immediately after that event. Pr is also referred to as the person's *prior probability function* and $\Pr_E$ as her *posterior probability function*.

Van Fraassen's rule, which will here be called "EXPL," is like Bayes' rule except that it attributes a bonus for explanatory superiority. Where $\{H_i\}_{i\leqslant n}$ is a set of self-consistent, mutually exclusive, and jointly exhaustive hypotheses, a person's new probability for H_i immediately after receiving E is in accordance with EXPL if and only if

$$\Pr'(H_i) \;=\; \frac{\Pr(H_i)\,\Pr(E \mid H_i) + f(H_i, E)}{\sum_{j=1}^{n}\big(\Pr(H_j)\,\Pr(E \mid H_j) + f(H_j, E)\big)}, \tag{EXPL}$$

where Pr and Pr′ are the prior and posterior probability function, respectively, and f is a function assigning a bonus c ($c \geqslant 0$) to the hypothesis that best explains E and nothing to the other hypotheses.

Whereas EXPL gives all credit to the best explanation, one could plausibly consider rules that credit a number of hypotheses in proportion to how well they explain the data. So for instance, the best explanation might get most of the credit, but the second-best explanation might also get some credit, and might even get almost as much credit if it is almost as good, qua explanation, as the best explanation. One could also consider giving some credit to the third-best, the fourth-best, and so on, explanation, where then the credit attributed gets less and less, again most plausibly reflecting the explanation quality of each individual hypothesis. Indeed, if a hypothesis would make for a particularly poor explanation of the data, one could even assign it a malus point.

Taking this idea as a starting point, Douven (2017, 2019, 2020a, 2022; also Douven & Wenmackers, 2017) formulates probabilistic versions of abduction that can credit individual hypotheses separately, in accordance to their explanation quality. Specifically, the rules he proposes are instances of the following schema:

$$\Pr'(H_i) \;=\; \frac{\Pr(H_i)\,\Pr(E \mid H_i) + c\,\Pr(H_i)\,\Pr(E \mid H_i)\mathscr{M}(H_i, E)}{\sum_{j=1}^{n}\big(\Pr(H_j)\,\Pr(E \mid H_j) + c\,\Pr(H_j)\,\Pr(E \mid H_j)\mathscr{M}(H_j, E)\big)}, \tag{S}$$

where Pr and Pr′ are as before, $\mathscr{M}$ is a measure of explanation quality, and with again $c \geqslant 0$.

Note that, as stated here, the above schema as well as EXPL have Bayes' rule as a limiting case, viz., if c is set to 0. One could thus say that advocates of either schema are committed to Bayesian updating in cases in which no explanatory considerations are at play. Alternatively, one could require c to be strictly greater than 0, thereby leaving entirely open how to update one's probabilities when explanation plays no role.

In principle, $\mathscr{M}$ can be any measure of explanation quality. Douven (2022) considers two in particular, one building on Popper's (1959) work and the other on Good's (1960) work. According to Popper's measure, hypothesis H's power to explain evidence E is given by

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

while according to Good's measure it is given by

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

Douven uses these measures (to illustrate certain normative points about abduction, to be discussed in Sect. 4) because they had performed well in empirical research (Douven & Schupbach, 2015a, 2015b), not necessarily because he thinks they are "objectively best."

The easiest way to think of these rules is that they first update a hypothesis' probability according to Bayes' rule, calculate that hypothesis' explanatory goodness (or badness, as the case may be) according to one of the above measures, add (or subtract) a percentage of the hypothesis' probability in proportion to its explanatory goodness (or badness), and then, as a final step, renormalize.

The details matter less than the general observation that there *are* precise versions of abduction, for example, instances of EXPL or the schema of Douven (2017, 2019, 2020a, 2022), and possibly many others. But although these schemata help to address the concern of lacking precision, they do raise a concern of their own, at least for anyone wishing to maintain the rationality of abductive reasoning. The new concern is that there appear to be *many* versions of abduction, without an indication of which of those is the *right* one, the one to be followed in our reasoning.

Douven (2017, 2022) proposes not to see this as a concern but rather to embrace the thought that abduction is a general idea—that explanation has a role to play in confirmation—that not only *can* be articulated in a diversity of ways but that *has* to be articulated differently in different contexts of use. Exactly how to reason abductively depends on what the reasoner's goals are, on the environment in which she is pursuing those goals, as well as on her capacities. Indeed, if Foley (1993) and others are right that we sometimes reason qualitatively—in terms of what to (categorically) believe—and sometimes quantitatively—in terms of what probabilities to assign—there may be times when we rely on something like Kuipers' or Lipton's versions of abduction, referenced in the previous section, and also times when instead we rely on EXPL or a kindred probabilistic rule.

The proposal to understand abduction as a broad idea, requiring further fleshing out depending on context and user, takes its cue from work by Gigerenzer (2000, 2002), Elqayam (2011, 2012), Schurz and Hertwig (2019), and others, arguing for an *ecological* conception of rationality. This work suggests that we must be willing to abandon the classical idea that rational reasoning is a matter of following a small number of universally valid principles and to acknowledge that the ability to pick the right learning tools for each particular situation is an important aspect of what we generally think of as human intelligence. In light of this work, the thought that rationality may require us to use one precisification of abduction in some contexts, another in other contexts, and perhaps Bayes' rule in yet other contexts, makes a lot of sense.

Nevertheless, philosophers love generality, and so they may not be easily persuadable to let go of the one-size-fits-all solution that Bayesianism appears to offer. And then there are still the arguments that were mentioned in the introduction, which aim to show that any deviation from Bayesian reasoning leads to irrationality. Before turning to those, I discuss some evidence seemingly showing that, in quite ordinary learning situations, people tend to reason abductively, by taking explanatory factors into consideration in ways that lead them to violate Bayes' rule. At a minimum, that puts some pressure on those wanting to stick to Bayesianism, given that it would require the attribution of massive error in people's learning practices.[1]

[1]Bayesians may be quick to point out that it is long known that people violate Bayesian principles; see the next section. However, many Bayesians still want to maintain that, *by and large*, their view is descriptively adequate (e.g., Oaksford & Chater, 2007)—which becomes harder to maintain with every newly discovered violation.

3 Abductive reasoning: Empirical support

Bayes' rule as well as the probabilistic versions of abduction form the core of competing accounts of rational updating. It is not necessary for such accounts to be descriptively accurate to a tee. But they should be at least broadly predictive of how humans update their probabilities. If not, why think that these accounts have any bearing on *human* rationality, rather than being some highly idealized form of robot epistemology? So, how do these accounts hold up against the experimental results?

To start with Bayes' rule, it is to be stressed that much of what is commonly advertised as evidence for Bayesianism is unrelated to the question of how people update their probabilities upon the receipt of new information and concerns probabilistic reasoning more broadly, for instance, whether people's static assignments of probabilities are coherent, that is, whether people's (subjective) probabilities are truly probabilities in that they conform to the probability calculus, at least by and large (Oaksford & Chater, 2007). Whereas there are quite a number of known results supporting the thought that people do obey Bayesian prescriptions, at least approximately, there are also reports of stark violations of these prescriptions, most famously in the work of Kahneman and various of his collaborators (e.g., Kahneman, Slovic, & Tversky, 1982; Tversky & Kahneman, 1983). However, Bayesians have tried to explain away such violations as being due to people's reliance on error-prone heuristics, or on their confusing the concept of probability with that of confirmation (see Tentori, Crupi, & Russo, 2013, for discussion).

Support specifically for the descriptive adequacy of Bayesian *updating* is hard to find. Griffiths and Tenenbaum (2006) present participants with random samplings from a closed interval (e.g., a random person's age, or the length of a random couple's marriage) and then ask for the upper bound of that interval (e.g., the person's life span, or the total duration of the marriage). They show that their participants' responses are close to what one would expect them to be on the assumption that they updated on the random outcome of the sampling via Bayes' rule. Note, though, that in this setup, explanatory considerations nowhere enter the picture, meaning that the assumption that the participants updated via an instance of EXPL or of the other schema discussed in the previous section would lead to the same predictions.

Besides, there is older work on updating that also explicitly compared people's updates with what those updates should be according to Bayesianism and this work reported strong evidence *against* Bayes' rule (Edwards, 1968; Marks & Clarkson, 1972; Fischhoff & Lichtenstein, 1978; Schum & Martin, 1982). Particularly worth mentioning in this regard is the research reported in Phillips and Edwards (1966), which involved a so-called bookbags-and-poker-chips experiment. In this type of experiment, participants are being informed about the contents of two containers (e.g., bookbags or urns), where these containers hold two types of objects (e.g., black and red poker chips, or blue and green balls) in different ratios. For instance, they might be told that the bag composition is 70/30 versus 50/50. The experimenters then randomly draw a number of objects from one of the bags, without disclosing to participants which bag it is. Finally, the participants are shown the sample and asked for their probability that the sample comes from the 70/30 bag rather than from the 50/50 bag. Using this setup, and comparing their participants' probability estimates with the probabilities for the two bags given the sample that were mandated by Bayes' rule, Phillips and Edwards found significant discrepancies between the former and the latter.

Here too, no attempt was made to contrast Bayesian updating with updating via some form of abduction. More recently, however, Douven and Schupbach (2015a) relied on basically the same experimental paradigm with the explicit aim of investigating whether deviations from Bayes' rule in participants' probability updates—if any deviations were found—could be due to the partici-

pants' taking into account explanatory considerations. These authors were specifically interested in three questions, to wit, first, how Bayesianism and explanationism—the normative view that people ought to reason abductively when explanatory factors are at play—compare in terms of descriptive adequacy; second, whether if judgments of explanatory goodness are found to have an essential role in updating, probabilities still play an important role, too, in updating; and third, what kind of explanatory judgments figure in updating, if any do.

To answer these questions, Douven and Schupbach slightly extended Phillips and Edwards' bookbags-and-pokerchips paradigm, the extension consisting of the additional gathering of judgments of explanation goodness, alongside that of probability judgments. Specifically, the procedure was as follows: Participants were interviewed individually and were, at the start, presented with two urns, labeled "urn A" and "urn B." They were shown that each urn contained forty balls, the composition being thirty black balls and ten white ones for urn A, and fifteen black balls and twenty-five white ones for urn B. Participants could consult this information at any time during the interview. Then the experimenter flipped a coin and, depending on the outcome, chose one or the other urn, outside of the participants' view. Next, from the chosen urn ten balls were drawn, one after the other, and without replacement. The balls were lined up before the participant as they were drawn. After each draw, the participant was asked the following questions:

(i) How well, in your opinion, does the hypothesis that urn A was selected explain the results from the drawings so far?
(ii) How well, in your opinion, does the hypothesis that urn B was selected explain those results?
(iii) How probable is it, in your opinion, that urn A was selected, given the results so far?

The two questions about explanatory goodness had to be answered by making a mark on a continuous scale with "extremely poor explanation" and "extremely good explanation" as anchors.

In their analysis, Douven and Schupbach fitted a number of linear regression models (in fact, so-called linear mixed-effects models; see Douven, 2022, for some background), each of which had the collected responses to question (iii) as dependent variable and at least the objective conditional probabilities that could be calculated for each participant and each drawing as predictor variable. The models differed in their further predictors. In one model there were *no* further predictors beyond objective conditional probabilities. A second one included as further predictors both the collected responses to question (i) and the collected responses to question (ii). A third one, finally, had besides objective conditional probabilities as a predictor also the computed *differences* between the participants' responses to question (i) and question (ii).

Adding predictors to a model tends to yield a model with better fit. Therefore, Douven and Schupbach compared the aforementioned models using the Akaike information criterion (AIC) and the Bayesian information criterion (BIC), which weigh model fit against model complexity. On both criteria, the third model, with objective probabilities and difference in judged explanatory goodness as predictors, did best, followed by the second model, and with the first model—the one with only objective probabilities as predictor—coming in at a very distant third place.

This result casts doubt on the claim that people update via Bayes' rule rather than via some version of abduction. If they updated via Bayes' rule, the smallest model should have come out on top.[2] What should make Douven and Schupbach's (2015a) result particularly unsettling for Bayesians is that not only were the participants' responses out of sync with Bayesian prescriptions (which could perhaps be explained away in terms of noise), but these deviations could be successfully accounted

[2] At least that is so assuming the so-called Principal Principle, according to which (roughly) subjective probabilities should equal objective probabilities if the latter are known. But this principle is also almost generally endorsed in the Bayesian community.

for in terms of the participants' giving weight to explanatory considerations. This is evidence that people factor in judgments of explanation quality when they update, at least in some contexts, and that they do so in a way that is essentially non-Bayesian.

It does *not* follow from Douven and Schupbach's analysis that explanatory considerations have any *systematic* impact on people's updates. In particular, it does not follow that people are following something like a probabilistic rule of abduction. This observation motivated follow-up research specifically directed at the question left open by Douven and Schupbach (2015a). In this research, Douven and Schupbach (2015b) used the objective probabilities from their earlier study in conjunction with Popper's and Good's measures of explanatory power to compute, separately for each participant and each draw, the explanatory power of the hypotheses at play in the experiment reported in Douven and Schupbach (2015a)—that is, that urn A had been selected, and that instead urn B had been selected—and then used the results of those computations together with the objective conditional probabilities as predictor variables to again regress the updates from the participants in the experiment from Douven and Schupbach (2015a).[3] In other words, where the analysis of Douven and Schupbach (2015a) had used *subjective* judgments of explanatory goodness as a predictor, in their new analysis, these authors fitted models that had *computed* explanatory goodness values as a predictor. More exactly, one model had the values computed according to Popper's measure as a predictor and the other had the values computed according to Good's measure as a predictor, while both shared objective conditional probabilities as a predictor. It was found that both models did considerably better in terms of AIC and BIC values than the model with only objective conditional probabilities as a predictor. That is compelling evidence that, at least in some contexts, explanatory considerations do play a *systematic* role in people's probabilistic updates: the way they help shape those updates can be captured by formal measures of explanatory power.

It was previously mentioned that we may not always reason quantitatively, in terms of probabilities, but may also sometimes reason in terms of categorical beliefs, and that therefore qualitative versions of abduction may have psychological reality as well. Just as there is little empirical work on quantitative versions of abduction, there is little work on qualitative versions. To our knowledge, the only research directly concerned with a qualitative version of abduction—specifically, Lipton's, as stated in the previous section—is to be found in Douven and Mirabile (2018). Recall that Lipton's version stressed the importance of the best explanation being not only good enough but also being appreciably superior to the second-best explanation. To investigate the descriptive adequacy of this version, Douven and Mirabile focused on the following questions:

(1) Does the quality of an explanation predict people's willingness to accept that explanation, and is there a quality threshold such that an explanation must be above that threshold for people to infer to it?
(2) Will it make a difference to people's perception of the quality of an explanation if they are introduced to a rival explanation? Will that make a difference to their preparedness to accept the former?
(3) If people are given two rival explanations, does it matter to their preparedness to inter to the best of those how much the explanations differ, in terms of quality?

Douven and Mirabile (2018) describe three experiments designed to answer these questions.

The three experiments used materials deriving from six basic scenarios, each presenting a fact alongside one (in the first experiment) or two (in the second and third experiments) possible explanations of that fact. The explanations could vary in quality; where two explanations appeared,

[3] In fact, Douven and Schupbach (2015b) looked at some other measures of explanatory power as well, but these did significantly worse than Popper's and Good's measures.

the explanations could vary in quality independently of each other. The participants were asked to answer three questions: (i) whether they were willing to infer to one of the explanations (or to *the* explanation, where only one was given); (ii) how likely, in their judgment, the explanations were; and (iii) how good the explanations were, qua explanations. The participants in the third experiment also always had to indicate how confident they were in their answer to question (i).

In their analysis, Douven and Mirabile found that how highly a person rates the quality of an explanation accurately predicts how willing she is to infer to that explanation, and also that the perceived quality needs to be above a certain threshold (which differed somewhat among participants) before the person will make the inference. Importantly, whereas the probability a participant assigned to an explanation was also a good predictor of whether the participant was willing to infer to that explanation, perceived explanation quality was a significantly better predictor. Furthermore, Douven and Mirabile found that people's willingness to infer to an explanation is reliably affected by whether that explanation is presented on its own or is accompanied by a rival explanation, even though their judgment of the quality of an explanation is *not* affected by that. There was a further reliable effect of the *quality* of the rival explanation on people's preparedness to infer to the other explanation. Specifically, Douven and Mirabile found that when their participants were presented an explanation alongside a rival explanation that was more or less as good, the participants were reliably less inclined to infer to the former explanation, whereas the effect of introducing a rival explanation tended to be smaller when that rival was a clearly poorer explanation.[4] In summary, Douven and Mirabile found positive answers to question (1), the second part of question (2), and question (3), but a negative answer to the first part of question (2).

It was already known that, in some form or other, explanation is involved in various cognitive processes, such as categorization (e.g., Williams & Lombrozo, 2010, 2013; Edwards et al., 2019; Vasilyeva & Lombrozo, 2020), generalization (e.g., Lombrozo & Gwynne, 2014), and understanding (Keil, 2006; Legare & Lombrozo, 2014; Walker & Lombrozo, 2017). The studies discussed in this section are among the first to look specifically at the role of explanation in belief updating. The outcomes of these studies should at least for psychologists be reason to take abduction in the context of belief change more seriously than they have so far done. For example, it would be interesting to have more information about the degree to which abductive reasoning depends on context, and also to know more about the actual cognitive mechanisms underlying or involved in that type of reasoning. But whatever the outcomes of such (hopefully) future research, is there any reason for *philosophers* to care about it? Section 5 makes a case for a positive answer to this question. But Section 4 first discusses the two main arguments commonly taken to suggest that philosophers can safely ignore abduction.

4 Is abduction a recipe for disaster?

People are susceptible to all sorts of biased thinking. They have a strong tendency to give more weight to information that favors their views than to information that challenges those views (the so-called confirmation bias); they tend to discount older information in favor of more recent information (the recency bias); they easily neglect prior probabilities in their quantitative reasoning (the base rate fallacy); they often overestimate their own abilities (the Dunning–Kruger effect); and on and on. Hence, the finding that people systematically attend to explanatory factors when changing their beliefs, or their probabilities, is of little significance from a normative standpoint. After all, it could just be one more bias, one more unfortunate but hard to unlearn cognitive habit. If, at the

[4] Note that this finding is in line with Douven and Schupbach's (2015a) finding that their model with differences in judged explanatory goodness as a predictor, next to objective probabilities, did best.

end of the day, we had to acknowledge as much, that would at most be a *slight* additional blow to our self-esteem.

Philosophers have given two main arguments for the claim that abduction, if it has psychological reality, is a bias indeed, and a quite detrimental one at that. Before looking at these arguments in some detail, there are two remarks to be made. First, whereas abduction is nowadays almost generally derided, in the 1970s and 1980s it was almost equally generally considered a paradigmatically sound form of reasoning. McMullin (1992) referred to it as "the inference that makes science," and Boyd (1984, 1985) argued that because (on his analysis) abduction is central to scientific reasoning and the methods of philosophy should be continuous with those of science, abduction should be central to philosophy as well. That the appreciation of abduction changed so dramatically has everything to do with the meteoric rise of Bayesian philosophy of science and Bayesian epistemology, for reasons to be seen shortly.

Second, it is to be noted that the arguments to be considered in the following are strictly concerned with probabilistic versions of abduction. As said, there may well be situations in which we want to rely on Lipton's or Kuipers' or a similar qualitative version of abduction. The present author is not aware of any arguments against these. Naturally, hard-nosed Bayesians will regard the fact that these versions are phrased in terms of categorical rather than graded belief as disqualifying in itself. But the view that the two notions of belief are both to be taken seriously (rather than dismissing the categorical notion as somehow having no place in scientific philosophy) is increasingly popular, and much recent work in epistemology has looked at how (in Foley's, 1993, terms) the epistemology of beliefs and the epistemology of degrees of belief are connected (see, e.g., the papers in Douven, ed., 2021). Nevertheless, from here on, the focus will be on the probabilistic versions of abduction, on which most of the recent discussion about abduction has centered.

4.1 The dynamic Dutch book argument

According to the widely endorsed betting concept of probability, the degree to which you believe that your favorite football team will win its next match is the price in cents at which you are willing to take either side in a bet that pays $ 1 if indeed the team will win that match and nothing if it does not win. So suppose that you have no preference for selling that bet (you have to pay $ 1 dollar if the proposition turns out to be true) or for buying it (you receive the dollar if the proposition turns out to be true) for the price of ¢ 30. Then your probability for your favorite football team winning its next match equals 0.3.

Bayesians have relied on this concept to argue that any failure of our probabilities to accord with the axioms of probability—that is, for our subjective probabilities to be probabilities properly speaking—means you are in an irrational belief state. That is because—they argue—any such failure exposes us to a so-called Dutch book, which is the standard name in the literature for a bet or set of bets that guarantee a negative net pay-off. For instance, according to one of the axioms of probability theory, logical truths, like "A or not A" (with A any proposition), should be assigned a probability of 1. According to another axiom, the probability of a disjunction of mutually incompatible propositions should equal the sum of the probabilities assigned to the separate disjuncts. Now suppose you believe A to a degree of 0.4 and its negation to a degree of 0.7. Obviously, A and its negation are mutually incompatible, so you should believe their disjunction to a degree of 0.4 + 0.7 = 1.1. On the other hand, that disjunction is a logical truth, and so you should believe it to a degree of 1. You are clearly violating the axioms of probability theory. Here is how that can be exploited: I offer you for the price of ¢ 40 a bet on A that pays $ 1 dollar if A is true and nothing if A is false. Given the degree to which you believe A, you are willing to buy that bet. At the same time, I offer you for the price of ¢ 70 a bet that pays $ 1 if the negation of A is true (so if A is false) and

nothing otherwise. Again given your degrees of belief, you are willing to buy that bet. Exactly one of A and its negation will turn out to be true, so you can be sure to receive exactly $ 1 dollar from me. Note, however, that you have already paid me $ 1.1, meaning that, whatever the future brings, you will have a net loss of ¢ 10. Betting is risky—you can always lose money. What is different here, however, and what—Bayesians have argued—makes this an exhibit of your irrationality, is that *you could have seen the loss coming*. Not only that; you could have figured out how to avoid it, to wit, by making your probabilities for A and its negation align with the probability axioms.

The axioms of probability theory have nothing to say about how to *change* your probabilities in response to new evidence. Bayesians proposed Bayes' rule as an answer to that question but it was already seen that there are alternatives to that rule, even ones which are very similar to it except that they take explanatory factors into account, such as the instances of EXPL and S stated in Section 2. Bayesians have complemented the above Dutch book argument, which is *static* in that it only looks at probabilities held *at the same time* by a *dynamic* Dutch book argument, which looks at the development of a person's probabilities *over time*. In the typical presentation, someone who changes her degrees of belief by a non-Bayesian rule for belief change is offered a number of bets at different points in time. It is then argued that the bets will all appear fair to the person at the time they are offered to her but are, if she engages in all of them, guaranteed to make her lose money in the end.

To make this concrete, here is an example. Let is be given that a certain coin either is fair or has a perfect bias for heads (every toss lands heads) or has a perfect bias for tails (every toss lands tails). Suppose that, initially, each of these possibilities is equally likely. We are allowed to toss the coin, but first a bookie offers us two bets, one that pays $ 48 if the first two tosses do *not* both land heads, and one that pays $ 600 if the first two tosses do both land heads and, in addition, the third toss lands tails. In light of our prior probabilities, $ 28 and, respectively, $ 25 appear reasonable prices for these bets. (For instance, our prior that the first two tosses land heads equals 5/12 and so our prior that they do *not* both land heads equals 7/12, and 7/12 × 48 equals 28; similarly for the other bet.) Suppose the bookie agrees to sell the bets at these prices. Then so far we have spent $ 53.

Now let us start flipping the coin and update our probabilities as we watch the outcomes of the first two tosses. Suppose at least one of them comes up tails. Then we have won the first bet but lost the second one, which means that we receive $ 48 but still have a net loss of $ 5 (we payed $ 53 for the bets, after all). This would be unfortunate but nothing out of the ordinary: it is in the nature of betting that the bettor should be prepared to accept losses. But now suppose that the first two tosses do both come up heads. We have now lost the first bet but may still win the second one, which would allow us to pocket $ 600, thereby making a net profit of $ 547. However, before we toss the coin a third time, the bookie approaches us again and now instead of proposing to sell any bets proposes to buy one, viz., a bet that pays $ 600 if the third toss lands tails. For what price should we be willing to sell it?

Here it matters which rule we use to update our probabilities for the three hypotheses of interest (that the coin has a perfect bias for heads, that it is fair, and that it has a perfect bias for tails) on the outcomes of the first two tosses. Suppose we use EXPL, with an explanation bonus of 0.1. Then, as can be easily verified, our probability for the third toss landing tails will be 0.08. Thus, we are willing to sell the designated bet for $ 48 (eight hundredth of what the bet pays if the third toss lands tails). But notice now that, whatever the outcome of the third toss, we will have lost money. If the third toss does land tails, we will receive $ 600 but have to pay the same amount; in the other case, we will not have to pay anything but will also not receive anything. However, whereas we have spent $ 53 on the bets we bought, we have only made $ 48 on the bet we sold. In short, we have again lost $ 5. Thus, we are bound to lose $ 5 no matter what.

One equally easily verifies that updating via Bayes' rule would not have led to this result. For had we used that rule, our probability for the third toss landing tails after watching the first two landing heads would have been 0.1, so that we would only have been willing to sell the bet to the bookie had she offered to pay (at least) \$ 60. And if she *had* bought the bet for that price, we would have made a profit of \$ 7.

This is easily generalized to any update rule deviating from Bayes' rule, so in particular, to any instance of EXPL or S, or indeed to any other probabilistic version of abduction. What, according to Bayesians, makes this so damning to non-Bayesian update rules is that, again, the user can figure out herself, before deciding to update via a non-Bayesian rule, that the threat of engaging in bets that are bound to lose her money will always be lurking. From this, they conclude that non-Bayesian updating betokens irrationality. And so in particular, using some probabilistic version of abduction betokens irrationality.

4.2 Expected error minimization

The dynamic Dutch book defense of Bayes' rule, and the Dutch book approach to defending Bayesianism generally, has lost much of its erstwhile popularity. Most Bayesians have come to regard this approach as addressing the wrong sort of rationality. Being liable to Dutch books is a practical problem and may therefore indicate that we fall short of meeting standards of practical rationality, that is, the rationality concerned with our actions. But the debate about how to update our probabilities concerns a question of epistemic rationality, that is, a question of what we can rationally believe and to what degree.

Joyce (1998) was the first to point out this problem, and in the same paper he proposed an alternative to the Dutch book approach, one in terms of error minimization. Joyce was in effect only concerned with the "static" part of Bayesianism—the claim that rationality requires our subjective probabilities to be probabilities properly speaking—and not with updating. What he argued was, in essence, that any person whose epistemic state is not in accordance with the static norms of Bayesianism (i.e., whose subjective probabilities are not probabilities properly so called) falls short of realizing our epistemic goal, which Joyce understands in terms of inaccuracy minimization. That is to say, if a person's subjective probabilities are not formally probabilities, the person could improve the accuracy of her epistemic state just be bringing her subjective probabilities in line with the formal requirements of probability theory.

There is a variety of ways to measure the accuracy of subjective probabilities, but by far the most popular one is the so-called Brier scoring rule. To explain this rule, let $\{H_i\}_{i=1}^n$ be a set of self-consistent, mutually exclusive, and collectively exhaustive hypotheses, and let δ_{ij} (for $i,j \in \{1,\ldots,n\}$) equal 1 if $i = j$ and equal 0 otherwise. Then, where H_j is the true hypothesis, a person who assigns subjective probabilities $\mathbf{p} = (p_1,\ldots,p_n)$ to the members of $\{H_i\}_{i=1}^n$, with p_i her probability for H_i, incurs a Brier score of $1/n \sum_{i=1}^n (\delta_{ij} - p_i)^2$. By way of illustration, suppose H_1, H_2, and H_3 are self-consistent, mutually exclusive and jointly exhaustive, and your subjective probabilities for these hypotheses are 0.1, 0.5, and 0.5, respectively. Suppose that of these hypotheses H_2 is the truth. Then your Brier score equals $\left((0.1)^2 + (1 - 0.5)^2 + (0.5)^2\right)/3 = 0.17$. Because your subjective probabilities do not sum to 1, they are not probabilities in the formal sense. Suppose you bring them into alignment with the probability axioms by lowering your probability for H_3 from 0.5 to 0.4. Then your Brier score comes to equal $\left((0.1)^2 + (1 - 0.5)^2 + (0.4)^2\right)/3 = 0.14$. Naturally, this *could* be a coincidence, and it is certainly not true that any way of making your subjective probabilities accord with the probability axioms would lower your Brier score; for instance, if you lower your probability for H_2 from 0.5 to 0.4, that will bring your subjective probabilities in line with the probability axioms, but your Brier score would go up by more than 0.03. However, Joyce's point

is that whenever your subjective probabilities fail to obey the probability axioms, there is a way to lower your Brier score—and thus take a step toward realizing your epistemic goal—just by bringing your subjective probabilities in line with those axioms. *That*, and not protection against Dutch bookies, is why conformity with the probability axioms is a rationality requirement for subjective probability, or so Joyce argues.

While Joyce did not address the issue of the rationality of Bayes' rule, others argued that Joyce's general approach could also be used to show the rationality of Bayesian updating. Most notably, Leitgeb and Pettigrew (2010) sought to show that just as (according to Joyce) having subjective probabilities that are not really probabilities is sub-optimal from the perspective of realizing our epistemic goal, so is updating in ways that stray from Bayesian prescriptions sub-optimal from that perspective. More exactly, their claim is that updating via Bayes' rule is both necessary and sufficient for minimizing the expected inaccuracy of our post-update subjective probabilities, where the expectation is relative to our pre-update subjective probabilities, and where inaccuracy is again measured by the Brier score. To illustrate the idea, consider again hypotheses H_1, H_2, and H_3, to which we assign probabilities 0.1, 0.5, and 0.4, respectively. One piece of evidence relevant to these hypotheses that we might obtain is E. Exactly how it is relevant is specified by the following probability distribution:

$$\begin{array}{lll} \Pr(\{w_{H_1E}\}) = 0.01 & \Pr(\{w_{H_2E}\}) = 0.25 & \Pr(\{w_{H_3E}\}) = 0.1 \\ \Pr(\{w_{H_1\overline{E}}\}) = 0.09 & \Pr(\{w_{H_2\overline{E}}\}) = 0.25 & \Pr(\{w_{H_3\overline{E}}\}) = 0.3 \end{array}$$

Here, w_{XY} is the possible world in which both X and Y are true, and $\overline{E}$ designates the negation of E. Note that we can get the probabilities of the various hypotheses from this simply by summing the probabilities of the worlds in which they hold true. For example, we derive from the above that our probability for H_1 is equal to $\Pr(\{w_{H_1E}\}) + \Pr(\{w_{H_1\overline{E}}\}) = 0.01 + 0.09$, which indeed equals 0.1.

If we do obtain evidence E, then, according to Bayesians, we should update on that new information via Bayes' rule. As one easily verifies, this would lead us to assign the following subjective probabilities to the relevant possible worlds:

$$\begin{array}{lll} \Pr_E(\{w_{H_1E}\}) \approx 0.028 & \Pr_E(\{w_{H_2E}\}) \approx 0.694 & \Pr_E(\{w_{H_3E}\}) \approx 0.278 \\ \Pr_E(\{w_{H_1\overline{E}}\}) = 0.0 & \Pr_E(\{w_{H_2\overline{E}}\}) = 0.0 & \Pr_E(\{w_{H_3\overline{E}}\}) = 0.0 \end{array}$$

Right now, before receiving the evidence, what is our expectation for the Brier score we would incur if we updated on E? To calculate this, we consider what our score would be were H_1 to hold, we calculate what it would were H_2 to hold, calculate what it would be were H_3 to hold, and take the weighted average of the three scores, the weights being our probabilities for the worlds that still be possible after the update. This yields an expected Brier score of approximately 0.158. Remarkably, if we minimize

$$0.01\big((1-x)^2 + y^2 + z^2\big) + 0.25\big(x^2 + (1-y)^2 + z^2\big) + 0.1\big(x^2 + y^2 + (1-z)^2\big)$$

subject to the constraint that $x + y + z = 1$, we find a minimum of (approximately) 0.158, and equally remarkably, we find this minimum precisely at (0.028, 0.694, 0.278), which are our post-update probabilities for the remaining possible worlds.

If instead of Bayes' rule we use EXPL, again with a bonus of 0.1, to update on E, supposing we do receive that evidence, then that would lead us to assign different probabilities to those possible worlds. Suppose we find H_1 worthy of the explanation bonus. Then our probability assignment

would become

$$\begin{array}{lll} \Pr_E(\{w_{H_1E}\}) \approx 0.239 & \Pr_E(\{w_{H_2E}\}) \approx 0.543 & \Pr_E(\{w_{H_3E}\}) \approx 0.217 \\ \Pr_E(\{w_{H_1\overline{E}}\}) = 0.0 & \Pr_E(\{w_{H_2\overline{E}}\}) = 0.0 & \Pr_E(\{w_{H_3\overline{E}}\}) = 0.0. \end{array}$$

And we already know that, with those probabilities, we are *not* minimizing our expected inaccuracy. Indeed, in this case, our expected Brier penalty would be approximately 0.184, so greater than the penalty of 0.158 we would incur were we to use Bayes' rule. Leitgeb and Pettigrew (2010) show that nothing of this is a coincidence: any update rule that minimizes expected inaccuracy is equivalent to Bayes' rule.

The conclusion seems to be exactly parallel to the one Joyce drew from his argument: Bayes' rule is rational because using it is most conducive to our epistemic goal of inaccuracy minimization and not because it serves some practical goal (such as offering protection against dynamic Dutch bookies).

4.3 The end of abduction?

Both of the arguments discussed in the foregoing have done much to cement the popularity of Bayes' rule, the inaccuracy minimization argument currently being considered the more compelling of the two, for the reasons explained. At first blush, one could indeed wonder how these arguments, and certainly the second one, leave any room for doubt about the irrationality, or at least sub-optimality (and how could using a sub-optimal rule not be irrational if an optimal rule is available?), of any form of non-Bayesian updating, including probabilistic versions of abduction.

On closer inspection, however, the arguments leave much to be desired. Both have specific shortcomings, and they share a general one. I start with the shortcomings specific to each argument. As for the dynamic Dutch book argument, we already encountered the critique that it seems unrelated to what is or should be at issue, to wit, epistemic rationality. Moreover, some authors have questioned the betting concept of probability, or indeed the existence of any direct connection between probability and willingness to engage in bets, on which that argument, as any Dutch book argument, ultimately relies (e.g., Williamson, 1998). Finally, it has been argued that we should not think of update rules in isolation, but rather as parts of packages of further epistemic as well as decision-theoretic principles, and that there are such packages that include EXPL or a kindred version of abduction and that shield one from being exploited by dynamic Dutch bookies (Douven, 1999, 2022). Specifically, there are packages that will lead their users to deny that all bets offered by the bookie are fair, even though use of the package may also lead to violations of Bayes' rule.

As for the more specific problems facing the inaccuracy minimization argument, first note that it is *not* quite an extension of Joyce's argument to the dynamic case. According to Joyce, concordance with the probability axioms guarantees inaccuracy minimization, not *expected* inaccuracy minimization, which is what Leitgeb and Pettigrew claim obedience to Bayes' rule guarantees. In fact, the difference is a bit more subtle still, given that what Leitgeb and Pettigrew actually argue for is that obedience to Bayes' rule guarantees expected *next-step* inaccuracy minimization—so concerning the inaccuracy of our subjective probabilities immediately after the update—not expected inaccuracy minimization *tout court*. What this means is that they leave open the possibility that your expectation of how inaccurate your subjective probabilities will be at some point in the future is greater supposing you are committed to Bayes' rule than if you commit to some non-Bayesian update rule, like EXPL for instance. Not only that: they leave open the possibility that you will ultimately end up having more accurate subjective probabilities if you update via some non-Bayesian rule than if you update via Bayes' rule. Their argument could still be compelling if they had given

a reason to believe that we should only, or at least first and foremost, care about expected next-step inaccuracy minimization. But they have not, and pre-theoretically the claim appears rather implausible.[5]

But there is a more general point to be made, which pertains to both arguments, to wit, that still nothing follows about non-Bayesian belief change if there are monetary (as per the dynamic Dutch book argument) or epistemic (as per the inaccuracy minimization argument) costs attached to it. Few things in life are for free. There are costs attached to having dinner in a restaurant, but that does not prevent you from eating out: if you pick the right restaurant, you will find the meal you get there worth the money and will be happy to pay the bill. For some reason, Bayesians have never even bothered asking whether non-Bayesian updating could have any benefits compared to Bayesian updating. The next section addresses that question.

5 The case for abduction

To see how abduction can be preferable over Bayes' rule, all things considered, let us start by asking what one may want from an update rule. We gather evidence in the hope of arriving at the truth concerning some matter of interest. Sometimes, the evidence informs us immediately about the truth of that matter. Is Susan in her office? We may be in the position to simply have a look and see Susan in her office, which settles the matter to everyone's (but the skeptic's) satisfaction. But often the matter is not so easily decided. Why did the dinosaurs go extinct? Piecing together various bits of evidence, we may become inclined to think that it was due to environmental changes brought about by some catastrophic event, like the impact of an asteroid on earth. In cases like this, instead of simply observing the truth of the matter, we try to infer the truth from the evidence, the inference typically being uncertain to a degree. Bayesians and advocates of rules like EXPL or other probabilistic versions of abduction agree that the inferential mechanism at play is to be thought of as a rule that outputs new subjective probabilities on the basis of evidential input.

A number of desiderata for rules of this sort naturally flow from the idea which also underlies the inaccuracy minimization arguments just discussed, viz., that truth is the ultimate epistemic goal: all our epistemic efforts are geared toward becoming certain of things that are true that they are true and of things that are false that they are false. The most general desideratum is, of course, that we want update rules to be conducive to realizing this goal. More specific ones are suggested by attending to the most relevant dimensions along which update rules can vary with respect to their truth conduciveness. To begin with, we want such rules to be *reliable* in that they typically lead us to become more confident in truths and less confident in falsehoods, and the more so the more evidence we obtain. All else being equal, we prefer more reliable rules over less reliable rules. In practice—especially in scientific practice—it will often be difficult to arrive *exactly* at the truth and we may have to settle for getting *close* to the truth, or *close enough* for all practical purposes. All else being equal, we prefer an update rule that leads us to spread our confidence close to the truth over one that leads us to spread our confidence further away from the truth. A last important desideratum stems from the fact that, again in practice, we are frequently under some time pressure to arrive at the truth. That an update rule *eventually* will make us confident in the truth is not so helpful in situations in which, for instance, becoming confident in the truth, or just becoming more confident in the truth than in any of its false rivals, or becoming sufficiently confident in a hypothesis close enough to the truth, is a matter of life and death. So, all else being equal, we prefer

[5] Independently, the Brier score is not as obviously compelling as the proponents of the inaccuracy minimization arguments take it to be; see Douven (2020b, 2023). And there are possible alternatives relative to which versions of abduction, rather than Bayes' rule, minimize expected inaccuracy; see Douven (2022, Sect. 5.2).

an update rule that increases our confidence in the truth *rapidly* over one that does so more slowly.

Ideally, an update rule makes us reliably and rapidly gain high confidence in the truth and nothing but the truth. More realistically, we have to be prepared to make trade-offs. A rule that rapidly concentrates our confidence in some small area of the space of possibilities may do so at the expense of accuracy; it may quickly get us in the vicinity of the truth but then be quite slow in taking us exactly at the truth. Other rules may be quicker in bringing us exactly at the truth though it may take longer for them to gear up and therefore they may be actually slower in bringing us in the vicinity of the truth. Or, one rule may often quickly take us quite close to the truth though also often make us invest high confidence in falsehoods, whereas another rule moves us toward the truth more slowly but also more reliably.

We cannot say in general which trade-off or trade-offs we should be prepared to make and which we should not. In some circumstances, it may be of the utmost importance to be able to quickly concentrate our confidence in a smallish region of the space of possibilities—for instance, it may be important for a medical doctor to be 95 percent certain that a patient's systolic blood pressure is between 110 mmHg and 130 mmHg—but then be further relatively unimportant to concentrate our confidence even more (e.g., becoming highly confident that the systolic pressure is between 117 mmHg and 122 mmHg may have no further consequences for how the doctor will treat the patient). In other circumstances, it may be more important to estimate some given parameter with great accuracy but there may be no pressure to do so quickly. We can imagine how different update rules serve our purposes best in the different situations.

It was mentioned previously that philosophers are strongly inclined to aim at generality, even universality. Among other things, they have aimed to state rules of rational thinking and behavior that apply in each and every situation. In line with this tradition, Bayesians have tried to argue that Bayes' rule is the rational update rule in all contexts, under any circumstances, regardless of who is to use it. As was also mentioned, however, many researchers—especially in psychology—have recently warmed to the idea that rationality is a context- and even agent-dependent matter, an idea often going under the name of "ecological rationality." While the proponents of this conception of rationality disagree on details, they share the view that rationality is a matter of picking the right tool in relation to whatever one's goals and abilities happen to be in the context of use. And not only may different people have different goals or possess different abilities in the same context, one and the same person may have different goals or different abilities in different contexts of use.

To illustrate, consider the possibility that, in some domains, there is a strong correlation between explanatoriness and truth in the sense that hypotheses concerning matters in those domains that strike us as being explanatorily powerful have a tendency to be true; that could be a contingent fact about us in relation to the world we inhabit, or it could be due to the workings of some evolutionary mechanisms. In other domains, there might be no such correlation, or a much weaker one. In domains of the former type, we might be better off using a version of abduction rather than a rule (like Bayes' rule) that does not take explanatory factors into account. In domains of the latter type, it might be counterproductive to rely on any version of abduction.

This observation is the starting point for the defense of abduction to be found in Douven (2018, 2020a, 2022). Rather than seeking to show that abduction is *the* rational update rule, Douven demonstrates that there are realistic circumstances under which probabilistic versions of abduction outperform Bayes' rule in offering a better trade-off between speed and accuracy, that is, between how rapidly our confidence gets concentrated in a small region of the space of possibilities and how close that region is to where the truth is located in the space. Accordingly, in those circumstances, it would make more sense to use any of those versions of abduction than to use Bayes' rule.

The demonstration comes in the form of various computer simulations, pitting Bayes' rule

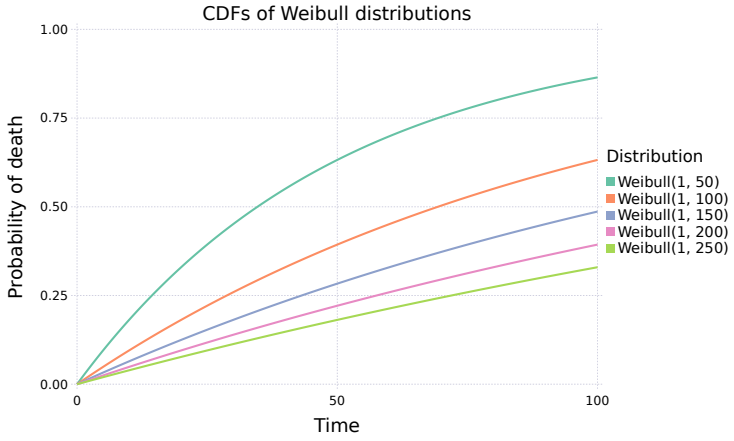

Figure 1: Examples of Weibull CDFs that give the probability of death of a patient as a function of time after admission into an intensive care unit.

and a number of different versions of abduction—instances of the schemata labeled "EXPL" and "S" in Section 2—against each other in contexts in which they are used to update sequentially on pieces of evidence received over time and related to some practical problem at issue. Here, one set of simulations will be described in detail and will also be generalized somewhat.

The simulations to be considered concern a setting in which medical doctors, working at an intensive care unit (ICU), are tasked to diagnose the patients who are brought into the unit and to determine, based on test results, how to treat the patient. Time is of the essence, given that the probability that the patient will die increases as time passes, though that probability decreases if the doctor makes the right intervention. By contrast, the probability that the patient will die *increases* if the doctor decides upon the wrong intervention.

How the probability of death increases with time, provided no intervention is made, can be modeled in various plausible ways. Douven (2020a, 2022) considers two options, one of which models this probability by the cumulative density function (CDF) of some Weibull distribution, and the other of which models that probability by the CDF of some Gamma distribution. Here, only the former will be described. Also, most of the formal details of Weibull distributions are skipped; it is only noted that they are characterized by a shape parameter and a scale parameter. Figure 1 shows five examples of a Weibull distribution, all having a shape parameter of 1 but having different scale parameters. In the simulations to be considered, the probability of death for a given patient brought into the ICU is assumed to be modeled by some Weibull distribution, where the shape parameter is, for each patient individually, chosen randomly and uniformly from the [0.5, 5] interval and the scale parameter is, also per patient, chosen randomly and uniformly from the [50, 250] interval.

In Douven's simulations, patients are further characterized by two parameters indicating how the right and, respectively, wrong intervention will impact the probability that the patient will die. Again skipping the formal details (for those, see Douven, 2020a, 2022), the idea is that making the right intervention lowers the probability of death by a certain percentage while making the wrong intervention increases that probability, the magnitude of the impact depending both on the patient and on the time of intervention. Figure 2 illustrates these effects for a specific parameter setting and a specific Weibull distribution.

Finally, what is wrong with a patient is, rather abstractly, taken to be a matter of the value a

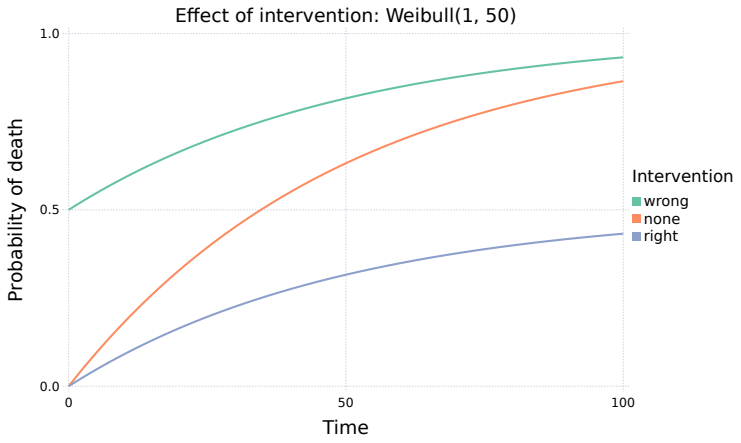

Figure 2: Examples of the effect of right and wrong interventions for a Weibull distribution, where the orange graph is the probability of death of the patient over time if no intervention is performed; the green graph gives, for every point in time, the probability of death of the patient if at that point in time a wrong intervention is performed; and the blue graph does the same for the correct intervention.

α assumes for her, the idea being that, as the patient enters the ICU, his medical status is know except for the value of this parameter. It is given, however, that this parameter can take a value in $\{0, .1, .2, \ldots, 1\}$ only, that the doctor knows this, and that she initially deems each of these values equally likely. The doctor receives one new test result per unit of time, on the basis of which she is to estimate the value of α, the results being either "positive" or "negative," and the tests being probabilistically independent of each other, with the same (unknown) probability of being positive. The hypothesis that $\alpha = x$ states that the probability for any given test turning up positive is x.

Doctors are fully characterized by the update rule they use to accommodate the test results. Some doctors are Bayesian updaters, others use an instance of EXPL, still other doctors use a version of "Popper's rule," which is an instance of S with $\mathscr{M}$ being Popper's measure of explanation quality, and yet other doctors use a version of "Good's rule," with $\mathscr{M}$ being Good's measure of explanation quality; in the case of the versions of abduction, different doctors can assume different explanation bonuses. The simulations assume that a doctor must be sufficiently certain about a hypothesis before she intervenes, where "sufficiently certain" was understood as having a subjective probability greater than 0.9 in the hypothesis. They further assume that a doctor will perform the correct intervention only if she becomes sufficiently certain about the true hypothesis; else, she will make an incorrect intervention, where it is stipulated that all incorrect interventions will have an equally big negative impact on the patient's survival chances.

The question the simulations then seek to answer is this: Given (as we may assume) that each doctor has the goal of saving her patients' lives, which update rule should she use to accommodate the test results she receives? Rather than just summarize the simulations reported in Douven's work, we would like to rerun them, adding a slight twist to them. The twist concerns the fact that, in Douven's simulations, there is a fixed decision threshold of 0.9. The choice of this value was not entirely arbitrary: in the literature on the connection between categorical belief (or acceptance) and subjective probability, many authors have proposed 0.9 as the threshold for belief. Needless to say, however, this is at best an idealization. It is more realistic to assume that different people have different (possibly context dependent) thresholds for belief. Indeed, in Douven's simulations, should the real question not have been which *combination* of update rule and decision threshold

serves best the doctors' shared goal of saving as many lives as possible?[6]

As Douven (2020a, 2022) explains, this question can be thought of as a constrained optimization problem, the constraint coming from the fact that our choice of update rules is limited to the ones mentioned previously. There appears to be no closed form of the objective function (i.e., the function to be optimized), due to which analytical methods are not going to be of much help in solving the problem. For that reason, Douven recruits a form of evolutionary computation, which is a well-known optimization technique. As the name suggests, this technique seeks to exploit the basic principles at work in the process of natural selection, where instead of organisms struggling for survival the units of selection are different solutions to a given problem, which can differ in their "fitness," the criterion of fitness being determined by the problem at hand. The algorithm starts by selecting from a pool of randomly generated solutions the "fittest" solutions to be retained and then typically lets the selected solutions "reproduce" in some specific way. The retained solutions together with their "children" form the pool for the next round of computations, in which the competition for survival and reproduction starts again. This is repeated either for a predetermined number of times or until a fixed point is reached at which all solutions are the same or at least are equally good (Barbati, Bruno, & Genovese, 2012).

As in the simulations documented in Douven (2020a, 2022), our procedure starts with a pool of 200 "medical doctors" (the first generation of solutions), with fifty doctors using Bayes' rule, fifty using an instance of EXPL, fifty using an instance of Good's rule, and fifty using an instance of Popper's rule. For all but the first of these groups, the value of the explanation bonus c is, for each doctor individually, chosen randomly and uniformly from the [0, 0.25] interval. In addition to what was done in Douven's simulations, where each doctor had the same fixed threshold for belief of 0.9, here a threshold value is picked for each doctor separately, where this value is chosen randomly and uniformly from the [0.5, 1] interval.[7] Each doctor treats one hundred patients, whose relevant characteristics (probability of survival, how that probability is affected by right and wrong interventions, and value of α) are chosen randomly and separately per patient, in the way specified previously.

The doctor can spend 100 units of time on the treatment of each patient, where at each moment, until the doctor decides to intervene (if at all), the doctor receives the outcome of a single test, which is positive with a probability determined by the value of α that was randomly picked for the agent. At start time, the doctor has the same subjective probability in all of the eleven hypotheses about the value of α. These probabilities are updated sequentially, as the test results come in, one per time step, and using the update rule associated with the doctor. As soon as the probability for one hypothesis exceeds the threshold associated with the given doctor, she intervenes. If that probability is assigned to the *true* hypothesis, the doctor receives a score determined by the probability of death associated with the *right* intervention at the time the probability crossed the threshold; if the doctor assigns a probability above the threshold to a *false* hypothesis, her score is determined by the probability of death associated with the *wrong* intervention at the time the probability crosses the threshold; and if *no* hypothesis is assigned a probability above the threshold during the 100 time steps, the doctor receives the score of 1 minus the probability of death at the 100-th time step.

After a doctor has treated 100 patients, her overall score is simply the mean of the scores received for each patient, which can be interpreted as the average patient survival rate for that doctor. Then the 100 "fittests" doctors—the doctors with the highest average patient survival rate—are selected to go on to the next generation, which they form together with a copy of themselves (so that this

[6] This question was raised independently by Paul Thorn and Zina Ward.

[7] It would make no sense to allow for values below 0.5, as such a value would mean that the doctor can believe things she deems less likely than their negation.

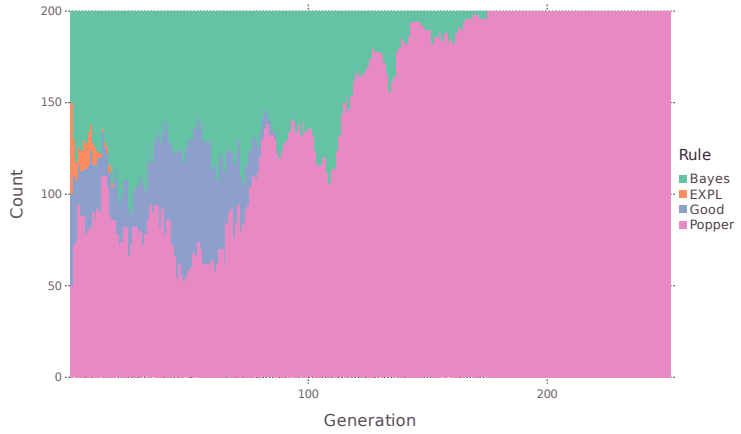

Figure 3: Counts of doctor types per generation for a randomly chosen simulation.

generation again consists of 200 doctors). This is repeated for 250 generations, after which the simulation terminates.

Fifty of these simulations were run. As an illustration, Figure 3 shows for one of those simulations how the pool of doctors evolved in the optimization process, with the generations represented on the x-axis and the count of doctors belonging to a certain group, characterized by the type of update rule they use, represented on the y-axis. It is seen that, in this simulation, Bayesians held up quite well for a while, but in the end the doctors updating via some instance of Popper's rule wiped out the competition entirely.

It is more informative to look at all simulations and consider the average number of doctors of the types at issue to be found in the 250 generations. These averages are shown in Figure 4. It is already somewhat clear from this figure that the simulation shown in Figure 3 is rather representative: Popperians were, overall, the clear winners, with Bayesians being a distant second. The first is very much in line with the findings reported in Douven (2020a, 2022), but Bayesians do in fact

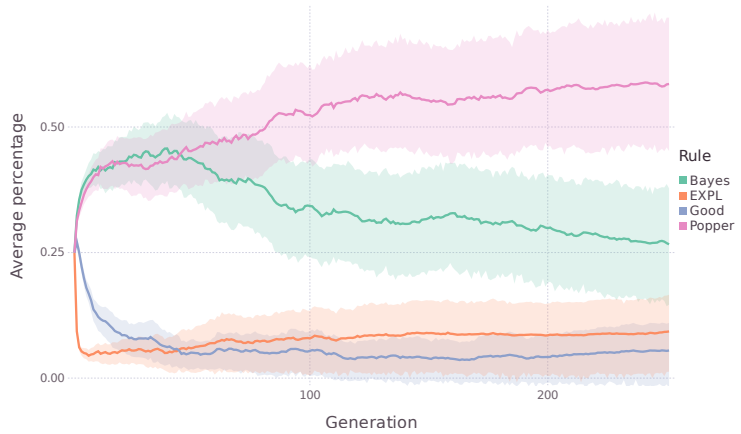

Figure 4: Percentages of doctor types per generation, averaged over the fifty simulations. Shaded areas indicate 95 percent confidence bands.

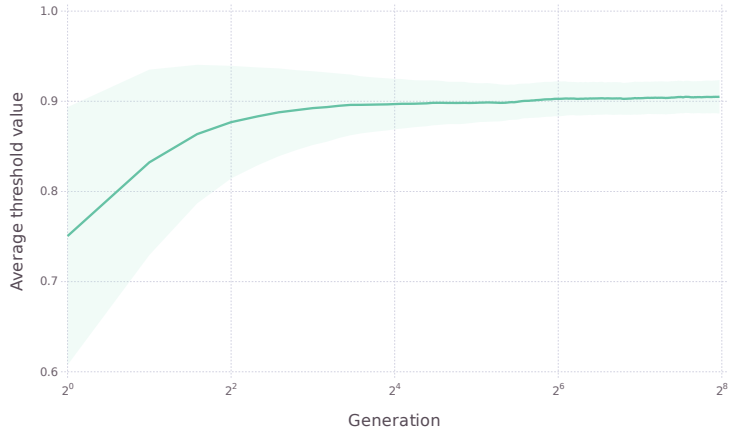

Figure 5: Log plot of average threshold value per generation, with 95 percent confidence band.

markedly better than in the previous simulations, where they ended up doing worse than the EXPL users, which is not the case here.

Hence, there *is* an effect of "unfixing" the threshold for intervention, albeit not one which changes our view that, in the present context, it is more advisable to use Popper's rule than any of the other rules, including Bayes' rule. Of course, threshold values were also subjected to evolutionary pressures in the new simulations. How did these impact them? The answer is highly surprising. As already seen in Figure 5, the mean threshold value converged to a value close to 0.9. To be precise, the average threshold value of the doctors in the last generation was 0.91 (± 0.02). As said, the choice of 0.9 as a threshold value in the previous simulations was not entirely arbitrary. However, it almost looks too good to be true that, when we make the threshold a parameter that can be optimized in the evolutionary process, we do find that this process drives this threshold to have an average of basically 0.9, with the vast majority of doctors having thresholds very close to that value. The author has been unable to find any bug in the code for the simulations that might account for this finding, though interested readers are invited to inspect the Julia code that was used for the simulations.[8]

In connection with these simulations, it is worth reiterating some of the observations already made in Douven (2020a, 2022). First, as noted there, whereas the evolutionary algorithm used in the simulations first and foremost serves as an optimization method, it can in the case at hand also be conceived as showing how evolution may have favored agents good at selecting the right update rule for the right environment. Second, a plausible explanation of why an explanation-based update rule is, in the context considered, preferable to Bayes' rule is that it allows for adaptive learning (by letting users increase or decrease the bonus for explanatory goodness), which Bayes' rule in itself does not do. Indeed, this point is reinforced by comparing the outcomes of the new simulations with those reported in Douven (2020a, 2022). In the former, Bayesians have acquired some flexibility—because of the flexible thresholds—that they did not have in the previous simulations.[9]

Most importantly, neither the new simulations nor the ones reported in Douven (2020a, 2022)

[8]The code is publicly available at this repository: https://github.com/IgorDouven/Abduction-Theory-and-Evidence.git.

[9]But then why do EXPL users and users of Good's rule do worse than Bayesians in the simulations, given that Bayesians still do not have as much flexibility as those other agents? As explained in Douven (2022), that has to do with the fact that EXPL users and users of Good's rule are unable to bring the explanation bonus quickly enough close enough to what the optimal value for that bonus would be for them. For Popperians, the bonus value is, on average, already from the beginning of the evolutionary process quite close to what the optimal value for them is.

aim to show that it is always more rational to update via abduction (in some form) than via Bayes' rule. Rather, their point is to help counter the claim made by Bayesians that it is *never* rational to update via abduction. Everything said in the foregoing is consistent with the insights of Elqayam, Gigerenzer, and others who have worked on ecological rationality, which imply that there is no one-size-fits-all norm of rationality and that instead different update rules may be called for in different contexts for different persons. In light of the work on ecological rationality, specifying a realistic type of situation in which we are better off by relying on a version of abduction is all a defense of this type of reasoning requires.

Finally, much ink has been spilled over the question of whether abduction is compatible with Bayesianism. The foregoing suggests that the answer is a resounding *yes* if we are willing to let go of the imperialist ideas that have traditionally accompanied defenses of both Bayes' rule and abduction. As was shown, there can be contexts in which abduction trumps Bayesian updating in all respects that matter in that context. But it is by no means ruled out that there are contexts in which one is better off using Bayes' rule. So it is not only the case that Bayesians and explanationists can be friends (as Lipton, 2004, Ch. 7, argues); we can all in good conscience *be* Bayesians and explanationists, just not at the same time.

6 Conclusion

The evidence showing that explanatory considerations play a role in how people adapt their subjective probabilities on the receipt of new information is not necessarily evidence that people are irrational. The arguments purporting to show otherwise—the dynamic Dutch book argument and the expected inaccuracy minimization argument—were defused. What is most fundamentally wrong with these arguments is that they only look at costs and not at possible benefits that might be worth the costs (granting that the costs are real, which we are under no obligation to do). Starting from an ecological conception of rationality, it was possible to go beyond defusing the criticisms leveled at abduction and to make a positive case for this mode of reasoning. The conclusion is not that abduction is a universally rational mode of reasoning, but rather that there are situations in which rationality recommends its use, leaving open the possibility that there are other situations in which one does better to rely on some other form of reasoning.[10]

References

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.

Boyd, R. N. (1984). On the current status of scientific realism. *Erkenntnis* 19: 45–90.

Boyd, R. N. (1985). Lex orandi est lex credendi. In P. Churchland & C. Hooker (eds.), *Images of science* (pp. 3–34). Chicago, IL: University of Chicago Press.

Douven, I. (1999). Inference to the best explanation made coherent. *Philosophy of Science* 66: S424–S435.

Douven, I. (2017). Inference to the best explanation: What is it? And why should we care? In K. McCain & T. Poston (eds.), *Best explanations: New essays on inference to the best explanation* (pp. 4–22). Oxford, UK: Oxford University Press.

Douven, I. (2019). Optimizing group learning: An evolutionary computing approach. *Artificial Intelligence* 275: 235–251.

[10] I am grateful to two anonymous referees for helpful comments on a previous version of this paper.

Douven, I. (2020a). The ecological rationality of explanatory reasoning. *Studies in History and Philosophy of Science* 79: 1–14.
Douven, I. (2020b). Scoring in context. *Synthese* 197: 1565–1580.
Douven, I. (ed.) (2021). *Lotteries, knowledge, and rational belief: Essays on the lottery paradox*. Cambridge, UK: Cambridge University Press.
Douven, I. (2022). *The art of abduction*. Cambridge, MA: MIT Press.
Douven, I. (2023). Scoring, context, and value. *Synthese*, in press.
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* 44: 1792–1813.
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, https://doi.org/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.
Edwards, B. J., Williams, J. J., Gentner, D. & Lombrozo, T. (2019). Explanation recruits comparison in a category-learning task. *Cognition* 185: 21–38.
Edwards, W. (1968). Conservatism in human information processing. In B. Kleinmuntz (ed.), *Formal representation of human judgment* (pp. 17–52). New York, NY: 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.
Fischhoff, B. & Lichtenstein, S. (1978). Don't attribute this to Reverend Bayes. *Psychological Bulletin* 85: 239–243.
Foley, R. (1993). *Working without a net*. Oxford, UK: Oxford University Press.
Gigerenzer, G. (2000). *Adaptive thinking: Rationality in the real world*. New York, NY: Oxford University Press.
Gigerenzer, G. (2001). The adaptive toolbox. In G. Gigerenzer & R. Selten (eds.), *Bounded rationality: The adaptive toolbox* (pp. 37–50). Cambridge, MA: MIT Press.
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.
Griffiths, T. L. & Tenenbaum, J. B. (2006). Optimal predictions in everyday cognition. *Psychological Science* 17: 767–773.
Joyce, J. (1998). A nonpragmatic vindication of probabilism. *Philosophy of Science* 65: 575–603.
Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge, UK: Cambridge University Press.
Keil, F. C. (2006). Explanation and understanding. *Annual Review of Psychology* 57: 227–254.
Kuipers, T. A. F. (1992). Naive and refined truth approximation. *Synthese* 93: 299–341.
Legare, C. H. & Lombrozo, T. (2014). Selective effects of explanation on learning during early childhood. *Journal of Experimental Child Psychology* 126: 198–212.
Leitgeb, H. & Pettigrew, R. (2010). An objective justification of Bayesianism II: The consequences of minimizing inaccuracy. *Philosophy of Science* 77: 236–272.
Lipton, P. (1993). Is the best good enough? *Proceedings of the Aristotelian Society* 93: 89–104.
Lipton, P. (2004). *Inference to the best explanation* (2nd ed.). London, UK: Routledge.

Lombrozo, T. & Gwynne, N. Z. (2014). Explanation and inference: Mechanistic and functional explanations guide property generalization. *Frontiers in Human Neuroscience* 8, https://doi.org/10.3389/fnhum.2014.00700.

Marks, D. F. & Clarkson, J. K. (1972). An explanation of conservatism in the bookbag-and-pokerchips situation. *Acta Psychologica* 36: 145–160.

McMullin, E. (1992). *The inference that makes science*. Milwaukee, WI: Marquette University Press.

Oaksford, M. & Chater, N. (2007). *Bayesian rationality*. Oxford, UK: 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, UK: Hutchinson.

Psillos, S. (2004). Inference to the best explanation and Bayesianism. In F. Stadler (ed.), *Induction and deduction in the sciences* (pp. 83–91). Dordrecht, Netherlands: Kluwer.

Schum, D. A. & Martin, A. W. (1982). Formal and empirical research on cascaded inference in jurisprudence. *Law and Society Review* 17: 105–151.

Schurz, G. & Hertwig, R. (2019). Cognitive success: A consequentialist account of rationality and cognition. *Topics in Cognitive Science* 11: 7–36.

Tentori, K., Crupi, V., & Russo, S. (2013). On the determinants of the conjunction fallacy: Probability versus inductive confirmation. *Journal of Experimental Psychology: General* 142: 235–255.

Tversky, A. & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review* 90: 293–315.

van Fraassen, B. C. (1989). *Laws and Symmetry*. Oxford: Oxford University Press.

Vasilyeva, N. & Lombrozo, T. (2020). Structural thinking about social categories: Evidence from formal explanations, generics, and generalization. *Cognition* 204: 104383.

Walker, C. M. & Lombrozo, T. (2017). Explaining the moral of the story. *Cognition* 167: 266–281.

Williams, J. J. & Lombrozo, T. (2013). Explanation and prior knowledge interact to guide learning. *Cognitive Psychology* 66: 55–84.

Williamson, T. (1998). Conditionalizing on knowledge. *British Journal for the Philosophy of Science* 49: 89–121.