3

The Psychology of Abduction

Without data, you're just another person with an opinion.

—W. Edwards Deming

3.1 Into the Mud

Lipton, whose work on abduction has done much to shape the current discussion, viewed that work as providing "a partial answer to the descriptive problem . . . of giving a principled account of the way we actually go about making non-demonstrative inferences" (Lipton, 2004, p. 142). So far, we have only pointed at some anecdotal evidence suggesting that abduction is not some exotic form of reasoning of strictly academic interest but rather captures an important part of how "we actually go about making non-demonstrative inferences." By today's standards, even in philosophy (in which empirically informed approaches have been gaining in popularity), I do not think that there is a way of tackling the descriptive question of how exactly we reason abductively without having recourse to the experimental techniques developed in cognitive psychology. In this chapter, we go beyond anecdotal evidence by giving a close look at data that directly bears on abductive reasoning. The data reveals in some detail what form this reasoning takes in certain well-delineated contexts. Part of the data subsequently inspires the statement of precisifications of abduction akin to EXPL.

The first pages of this book briefly discuss research, some of it going back to Helmholtz, about how explanatory considerations may shape what we perceive, even if only subconsciously. Research into how explanation connects to higher-level cognitive processes, such as learning, categorization, and

reasoning—research that has been only cited—is of a much more recent date. In retrospect, it is surprising that psychologists did not become interested in explanation earlier, because at least some of that research shows explanation to play a major role in cognition. For instance, Michelene Chi and colleagues (1994) show that prompting students to self-explain a text about the circulatory system led to better results on a later test than reading the study materials twice; this was so even if the self-explanations were misguided.[1] Furthermore, Joseph Williams and Tania Lombrozo (2010) have presented experimental findings showing that participants who were asked to *explain* why items belonged to a certain category were much more likely to discover the rules underlying the categorization than participants who were asked to *describe* the items (but who were also tasked to discover the categorization rule). There is, however, experimental work showing that the influence of explanation on cognition is not always for the good, in that, for example, it leads us to overestimate probabilities (Koehler, 1991) or it lets us take into account base rates that are evidentially irrelevant (Johnson, Rajeev-Kumar, & Keil, 2016).

In this chapter, we look at experimental results more directly concerned with the connection between explanation and belief formation or belief change—which is what abduction is about. Most of that research is of a very recent date. To appreciate the findings to be discussed, it is important to know about the dominant paradigm in current reasoning research, the so-called New Paradigm in the psychology of reasoning (Over, 2009; Elqayam & Over, 2012; Elqayam, 2018; Oaksford & Chater, 2020b). We start by outlining the fundamentals of that paradigm, which includes stating in greater formal detail the basics of Bayesianism, on which the paradigm builds. We then look at experimental findings concerning belief and belief change that allow us to compare a strict Bayesian view with one that leaves room for the possibility of abductive reasoning, in deviation from strict Bayesian principles. An attempt to explain away the apparent influence of explanatory reasoning on belief change as a result of systematic bias is shown to fail. In fact, it turns out that this influence can benefit our reasoning in that it makes us more accurate.

1. See, in the same vein, Sidney, Hattikudur, and Alibali (2015), who report that when they prompted students to self-explain examples of fraction division, that improved those students' conceptual learning.

3.2 The New Paradigm and Bayesian Rationality

From its beginnings in the 1960s, most notably through the work of Peter Wason, the psychology of reasoning was committed to the view that the norms of rational reasoning were given by classical logic.[2] Researchers saw it as their task to determine the extent to which people obeyed those norms, and to see whether any deviations from the norms could be attributed to systematic biases. Wason's (1968) celebrated selection task was the first of many whose results were all interpreted as showing that, in general, people were poor at logical reasoning and subject to a variety of biases. In the selection task, participants are presented with four cards, which they are informed have a letter on one side and a number on the reverse. The cards are laid out on a table, one with A showing, one with K, one with 2, and one with 7. The participants are tasked to determine whether it holds as a rule that a card has a 2 on one side if it has an A on the other side. They are asked to select the cards that must be turned in order to find out whether the rule holds for the four cards. Logic suggests turning the cards that show the A and the 7, which are the only ones that could provide falsifying instances, but that is not what people typically do; instead, they typically select the cards showing the A and the 2 or the A card alone.

At first, psychologists tried to explain these and similar apparent violations of logic in terms of cognitive limitations. The idea was that people *are* rational, at least in principle, but due notably to our limited working memory, we make performance errors when going through the necessary logical steps. Not much later, however, a number of psychologists started to question the logic-oriented approach generally. Logic was developed to facilitate mathematical reasoning, in which we use truth-preserving inference rules to derive theorems from a typically small set of supposedly self-evident axioms. Although logic has been immensely useful, it is to be acknowledged that much of our reasoning is not of that sort. Recall Schurz and Hertwig's (2019) previously cited point that the maximum ecological validity of logical

2. Wason was influenced by the work of the developmental psychologists Inhelder and Piaget (see in particular their 1958 book). But while Inhelder and Piaget held that logic was also a descriptively accurate model of how adults reason, Wason saw it only as a normative model and indeed took his work on the selection task described in the main text to have refuted the descriptive claim. In the terminology of Elqayam and Evans (2011), Wason was willing to commit only to *prescriptive* logicism, not to *descriptive* logicism; Inhelder and Piaget endorsed both.

inferences—in any context, if you use the rules of logic to derive a conclusion from a set of true premises, that conclusion is guaranteed to be true—is offset by their low applicability.

More specifically with respect to the selection task, researchers began to realize that participants may have differing degrees of confidence in the propositions involved in a reasoning experiment, and that those degrees of confidence and not just the logical form of the problem with which they are presented may impact their responses. For instance, experimenters interested in whether people reason in accordance with some argument form that is logically valid may ask their participants to accept certain premises and then ask to what degree they are willing to accept some conclusion, where the premises and conclusion together instantiate the argument form at issue. As mentioned, the suspicion was that the experimenters' request to assume the premises does not always suffice to completely shield against an impact on the participants' responses of their degrees of confidence in the premises. Numerous experiments confirmed this suspicion to be correct (Oaksford & Chater, 2001). In fact, it turned out that not only participants' degrees of confidence in the premises can tilt the results but their confidence in the conclusion can do so as well: various experiments showed participants to be more inclined to accept the conclusion of an argument if they already believed that conclusion to be likely, independent of the form of the argument; this effect is known as "belief bias" (see Evans, Barston, & Pollard, 1983).

Guided by such observations, psychologists came to regard degrees of confidence, or degrees of belief, as central to human reasoning. Concomitantly, they came to think that the norms of rationality should primarily pertain to such degrees of confidence—about which classical logic has little to say (but not nothing, as subsequently shown). Instead, taking their cue from work by Frank Ramsey (e.g., 1926) and Bruno de Finetti (e.g., 1937), and also influenced by more recent developments in philosophy and elsewhere, psychologists endorsed what is now generally known as "Bayesianism" as an encompassing account of rationality for graded beliefs.

3.2.1 Bayesian Basics

De Finetti, Ramsey, and others contributing to the development of Bayesianism took for granted that our beliefs come in degrees. That should not raise any eyebrows, as it is introspectively clear that we believe some things more

strongly or to a higher degree than others; we do not need empirical evidence for that (even though by now there is a wealth of supporting evidence).[3] Then the following are presented as the core principles of rationality:

Synchronic coherence (SC) At any time, our degrees of belief should conform to the probability calculus.

Diachronic coherence (DC) We should adapt our degrees of belief to the receipt of new information by means of Bayes's rule.

SC refers to the probability calculus, and DC to Bayes's rule; let us be specific about these.

Where "$\bot$" stands for an arbitrary contradiction, "$\top$" for an arbitrary tautology, and "$\vdash$" for the relation of logical consequence, a probability function is any real-valued function Pr that satisfies the following axioms:[4]

Axiom 3.1 For all propositions φ, $\Pr(\bot) = 0 \leqslant \Pr(\varphi) \leqslant 1 = \Pr(\top)$.

Axiom 3.2 For all propositions φ and ψ, if $\varphi \vdash \neg\psi$, then $\Pr(\varphi \vee \psi) = \Pr(\varphi) + \Pr(\psi)$.

In words, probabilities lie in the interval [0, 1]; contradictions always receive probability 0; tautologies always receive probability 1; and whenever two

3. Note, however, that for de Finetti and Ramsey, who were both operationalists, introspection had no value. Instead, they identified a person's degree of belief in a given proposition with her fair betting quotient for that proposition, that is, the price at which she is willing to take either side of a bet that pays off a certain positive amount of money if the proposition turns out true, and nothing otherwise. For instance, if a person is indifferent between buying and selling a bet on φ that pays y dollars if this bet's price is x dollars, then, it is said, she believes φ to a degree of x/y. Thus, if you are willing to take either side of a bet that pays $1,000 if it rains tomorrow (and nothing otherwise) if its price is $750, then it is said that you believe the proposition that it will rain tomorrow to a degree of 0.75. Because, de Finetti and Ramsey thought, we can always make measurements in this way, it always makes sense to attribute degrees of belief to people. From a modern psychological perspective, however, it is perfectly legitimate to cite introspective evidence for the existence of degrees of belief (Douven & Uffink, 2003).

4. Standard probability theory builds on *classical* logic; so, for instance, "$\vdash$" is taken to designate the relation of *classical* logical consequence. For attempts to define probability in a way that is neutral regarding the underlying logic, see Weatherson (2003); see Dietz (2010) for critical discussion.

propositions exclude one another, the probability of their disjunction is the sum of their separate probabilities.[5]

It is common practice in presentations of probability theory to define a shorthand notation for one especially important concept, to wit, that of conditional probability:

$$\Pr(\psi \mid \varphi) \quad =_{\text{df}} \quad \frac{\Pr(\varphi \wedge \psi)}{\Pr(\varphi)}, \quad \text{provided} \quad \Pr(\varphi) > 0.$$

We read $\Pr(\psi \mid \varphi)$ as the probability of ψ given (or "conditional on") φ.

Furthermore, Bayesians say that φ *confirms* ψ precisely if $\Pr(\psi \mid \varphi) > \Pr(\psi)$, *disconfirms* ψ precisely if $\Pr(\psi \mid \varphi) < \Pr(\psi)$, and is *neutral* with regard to ψ otherwise. There is an ongoing debate about how to define *strength* of confirmation. One of the earliest proposals (Carnap, 1950) is that φ confirms ψ (if at all) to a degree of $\Pr(\psi \mid \varphi) - \Pr(\psi)$, but many other measures have been proposed since (for discussion, see Fitelson, 1999, 2001; Fitelson & Eells, 2000; Eells & Fitelson, 2002; Crupi, Tentori, & Gonzalez, 2007; Brössel, 2013; Douven, 2021).

From axioms 3.1 and 3.2 and the definition of conditional probability, one readily derives, for instance,

Proposition 3.1 *For all φ and ψ:*

1. $\Pr(\varphi) = 1 - \Pr(\neg\varphi)$*;*
2. $\Pr(\varphi) \leqslant \Pr(\psi)$ *whenever* $\varphi \vdash \psi$*;*
3. $\Pr(\varphi) = \Pr(\psi)$ *whenever* $\varphi \equiv \psi$*;*
4. $\Pr(\varphi) = \Pr(\psi)\Pr(\varphi \mid \psi) + \Pr(\neg\varphi)\Pr(\varphi \mid \neg\psi)$ *whenever* $0 < \Pr(\psi) < 1$*;*
5. $\Pr(\psi \mid \varphi) = 1$ *whenever* $\varphi \vdash \psi$ *and* $\Pr(\varphi) > 0$*;*
6. $\Pr(\psi \mid \varphi) = \Pr(\psi)\big(\Pr(\varphi \mid \psi)/\Pr(\varphi)\big)$ *whenever* $\Pr(\varphi) > 0$.

While there is nothing deep about these consequences of probability theory, they are important from a psychological point of view inasmuch as it is

5. According to some Bayesians, axiom 3.1 should be strengthened so as to require that *only* contradictions receive probability 0 and *only* tautologies receive probability 1. Some also propose as a further axiom a "continuity" principle according to which, for any countably infinite inconsistent set $\{\varphi_1, \varphi_2, \varphi_3, \ldots\}$ such that $\varphi_{n+1} \vdash \varphi_n$ for all n, it holds that $\lim_{n\to\infty} \Pr(\varphi_n) = 0$. However, for the purposes of this book, we can sidestep these and other more advanced technical issues. For a detailed yet accessible discussion of such issues, see Weisberg (2011).

straightforward to test whether people's degrees of belief are in accordance with them, which gives an easy opportunity to put descriptive Bayesianism (see below) to the test. The same is true for the Bayesian notion of confirmation and in particular for measures of confirmation (Crupi, Tentori, & Gonzalez, 2007; Tentori et al., 2007).

Item 6 of proposition 3.1 is also known as "Bayes's theorem" and is of particular importance because it plays a key role in confirmation theory and therefore in what follows. To appreciate its central place in confirmation theory, let us first rewrite it naming the variables more suggestively:

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

Think of H as a hypothesis in which we are interested and of E as evidence that we gathered. Then what the theorem says is that we can determine the probability $\Pr(H \mid E)$ of that hypothesis in light of our new evidence on the basis of the likelihood H bestows upon E, $\Pr(E \mid H)$, and of how likely we deemed H and E to be before we obtained the evidence, given by $\Pr(H)$ and $\Pr(E)$, respectively. This is important because it is often not immediately clear what value $\Pr(H \mid E)$ should be assigned, whereas we often simply *know* the probabilities that we assigned to H and E prior to learning E. Determining the likelihood is often also not so hard. When H is a statistical hypothesis—which invariably is the case in the applications considered subsequently—the likelihood follows analytically.[6] For instance, if H says that a given coin is fair, then the likelihood that this hypothesis bestows on the evidence that a toss with the coin comes up heads is obviously .5; more generally, if H attributes a bias for heads of x to the coin, then $\Pr(\text{Heads} \mid H) = x$. Simpler still, if the hypothesis logically *implies* the evidence, then, by item 5 of proposition 3.1, the likelihood equals 1.

Bayes's theorem is distinct from Bayes's rule (also known as "conditionalization" or "conditioning"), which according to DC we should use to update our degrees of belief if we want to be rational. This rule says that if you have learned φ, in the sense that you have become subjectively certain of φ, and nothing stronger—nothing that logically entails φ but is not entailed by φ—then for all ψ your new degree of belief in ψ should equal the conditional degree of belief in ψ given φ that you had just before learning φ. That is, where

6. Or at least "analytically"; see again chapter 1, footnote 21.

Pr is your degrees-of-belief function *prior* to learning φ and $\Pr_\varphi$ is your new degrees-of-belief function immediately *after* learning φ, Bayes's rule requires that it hold for all ψ that $\Pr_\varphi(\psi) = \Pr(\psi \mid \varphi)$.

As said, Bayes's rule is intended to apply whenever we have learned something in the sense of having become *certain* of it. But, naturally, not all learning is like that. By briefly peeking into a badly lit room, we may become *more* certain that the color of the carpet in the room is orange even if it does not make us fully certain about that color or indeed about anything about which we were not already certain beforehand. We may receive statistical information, such as that smokers are at a higher risk of developing coronary artery disease than nonsmokers. Thereby we have clearly *learned* something, but it is not the kind of evidence to which Bayes's rule applies. Again, it has been argued that the rule does not apply when we come to learn conditional information, such as that London will be flooded if global warming continues (Douven, 2012b; see also Douven & Dietz, 2011). Bayesians have come up with different proposals for accommodating each of these kinds of evidence.

For the first kind, Richard Jeffrey (1983, ch. 11) proposed a generalization of Bayes's rule, which now commonly goes by the name "Jeffrey conditionalization," and which (roughly) makes one's new degree of belief in the target proposition (in our example, that the carpet is orange) a weighted average of one's conditional degrees of belief in the proposition given the various possible evidence propositions (perhaps the propositions that we saw an orange carpet, and its negation), where the new degrees of belief assigned to those evidence propositions after the learning experience serve as weights. James Joyce (1999, ch. 6) gives a useful overview of the variety of rules that have been proposed for accommodating learning events of the second type (see also Uffink, 1995). And Bayesian procedures for accommodating conditional information have been proposed by Mario Günther (2018) and Benjamin Eva, Stephan Hartmann, and Soroush Rafiee Rad (2020).[7]

Returning to the core principles, SC and DC: why should we accept them? The dynamic Dutch book argument and the inaccuracy-minimization argument have already been mentioned as the most commonly cited reasons for holding that Bayes's rule is the only rational update rule. Each of these arguments was in fact preceded by a version addressing SC: Ramsey and (independently) de Finetti were the first to argue that if and only if our degrees

7. See also Sprenger and Hartmann (2019, ch. 4).

of belief conform to the axioms of probability are we safe from Dutch bookies (Vineberg, 2016). And Joyce (1998) purported to show that degrees of belief that are not probabilities are not as accurate as they could otherwise be. In the following discussion, I am concerned only with the arguments in support of DC. What I have to say about these arguments does *not* imply that Bayes's rule is necessarily wrong. Rather, the claim is that Bayesians are mistaken when they hold that whenever our evidence is not of the uncertain kind for which Jeffrey conditionalization was devised, or consisting of statistical information, or conditional in form, Bayes's rule offers the only rational way to update one's degrees of belief. Sometimes the best way to update may be via Bayes's rule, but sometimes it is via some precisification of abduction and sometimes (perhaps) via some altogether different rule. So, to the extent that people are rational, we should not expect them to always follow Bayes's rule; in other words, we should expect descriptive Bayesianism, to which we turn in the next section, to be false as a general hypothesis.

As a final comment, I note that in the way presented here Bayesianism is highly subjective, stemming from the fact that axioms 3.1 and 3.2 impose only very weak constraints on the function Pr. It could be the representation of *your* degrees-of-belief function, or of mine, or of anyone else's, provided that only these functions satisfy the said axioms. But those axioms leave a tremendous amount of room for variation, so much so that some believe them to lead to a debilitating form of relativism, a relativism according to which for some piece of evidence E and some hypothesis H, E might be evidence for H from your perspective, evidence against H from mine, and neutral with regard to H from a third person's perspective. As far as Bayesianism goes, we may all be "right." But we feel queasy about the idea that whereas we just obtained experimental evidence *against* a certain hypothesis, it could have been evidence *for* that hypothesis if only we had started from a different prior probability assignment.

Not all Bayesians share this sentiment. As for instance the Bayesian statistician Dennis Lindley writes,

> I am often asked if the [Bayesian] method gives the right answer, or, more particularly, how do you know if you have got the *right* prior. My reply is that I don't know what is meant by "right" in this context. The Bayesian theory is about coherence, not about right or wrong. (Lindley, 1976, p. 359)

To the eyes of many, however, this is little more than a restatement of the problem. It is unsurprising therefore, that various authors have taken to supplement Bayesianism with a number of further (supposedly) rationality requirements. Ideally, such further requirements allow us to make sense of the idea that there is (in Lindley's terms) a right prior, so that rational persons, provided they have the same background knowledge, would have to start from the same probability assignment. Somewhat less ideally, they rule out at least whole classes of priors as *not* right.

The Bayesian literature is rife with discussion of candidates for additional rationality requirements. One well-known proposal in this regard is Lewis's (1980) Principal Principle, according to which, roughly, one's degree of belief in φ given that the objective probability or chance of φ equals x, should equal x. But, given that for many propositions the notion of objective chance makes no sense (according to some, it makes sense only for propositions that are about quantum-mechanical events), this constraint is manifestly too weak to block the subjectivism previously pointed out. A seemingly more promising proposal has been the Principle of Indifference, mentioned in section 1.2.2. In its simplest version, this says that where $\{H_1, \ldots, H_n\}$ is a set of mutually exclusive hypotheses the disjunction of which is a logical truth, then in the absence of reasons to the contrary, you ought to assign each H_i the same probability, $1/n$. For example, if all we know about a given coin is that its bias for heads is a whole multiple of .1 (so it might be 0, .1, .2, and so on, until 1), then the said principle enjoins us to assign a probability of $1/11$ to each element of $\{H_1, \ldots, H_{11}\}$, where H_i is the hypothesis that the coin has a bias of $(i - 1) \times .1$ for heads.

As also mentioned, however, the principle may appear plausible at first, but on closer inspection it is hard to maintain. The problem is that it imposes inconsistent requirements. To see this, consider the set $S = \{H_1, \ldots, H_n\}$ of mutually exclusive hypotheses whose disjunction is a logical truth, and assume $n \geqslant 3$. Then $S' = \{H_1, \bigvee_{i=2}^{n} H_i\}$ is also a set of mutually exclusive hypotheses whose disjunction is a logical truth. It is perfectly consistent to assume both that you lack reasons to favor any hypothesis in S and that you lack reasons to favor either hypothesis in S'. However, given the former assumption, the Principle of Indifference enjoins you to set $\Pr(H_1) = 1/n$, whereas, given the latter, it enjoins you to set $\Pr(H_1) = 1/2$, which you cannot consistently do.

We looked at some ideas about how an appeal to explanatory considerations might help to save the principle but found those wanting, too.[8]

3.2.2 Descriptive Bayesianism

Descriptive Bayesianism is the view that SC and DC are not *merely* norms—which people might fail to satisfy by a wide margin—but are also *actually* obeyed by people, or at least that people behave cognitively *as if* their aim was to obey these norms, by and large. "By and large" is an important qualification. According to axiom 3.1, we should believe every logical falsehood to a degree of 0 and every logical truth to a degree of 1, no matter how difficult it may be for ordinary mortals to detect that they are logically false or logically true.[9] No psychologist expects people to fully satisfy this requirement or regards a failure to satisfy the requirement as a token of irrationality. Similarly, whereas item 2 of proposition 3.1 appears to require that if φ entails ψ, we assign a probability to the latter at least as high as the probability we assign to the former, this can translate to a requirement of rationality only with the qualification that we be able to know, or perhaps even easily detect, the holding of the entailment relation. Again, the same point could be made with respect to item 3 of proposition 3.1: if we cannot reasonably be expected to see the equivalence between two propositions, then we cannot reasonably be criticized for assigning different probabilities to them.

That we are not logically omniscient does not exclude our being rational. But assigning a probability of, for example, .7 to the possibility of rain tomorrow and of .5 to the possibility of no rain tomorrow does betoken irrationality, according to descriptive Bayesians. That thereby we are violating the axioms of probability requires no logical omniscience or other superpowers on our

8. Decock, Douven, and Sznajder (2016) have proposed a demonstrably consistent version of the Principle of Indifference, capitalizing on Gärdenfors's conceptual spaces approach (Gärdenfors, 2000, 2014; Douven & Gärdenfors, 2020). For all we presently know, however, only certain families of concepts are representable in conceptual spaces. If so, the aforementioned version of the Principle of Indifference has limited applicability.

9. As a matter of fact, given that Bayesians usually understand the $\vdash$-relation as stronger than logical entailment and as "incorporating all of contemporary mathematics" (Howson & Urbach, 1993, p. 20) and correspondingly take $\bot$ and $\top$ to designate not only logical but also mathematical falsehoods and truths, a strict descriptive Bayesianism would require us to assign probability 0 to all mathematical falsehoods and probability 1 to all mathematical truths, on pain of irrationality, no matter if even the greatest mathematicians could not (yet) recognize them as such.

part. Similarly when it is, or should be, obvious to a person that φ entails ψ and yet she believes the former to a higher degree than the latter, or if we know that the equivalence of the two propositions is transparent to her but she believes them to different degrees. Essentially the same point could be made with respect to many other straightforward consequences of probability theory. In other words, we can distinguish between interesting and uninteresting failures to obey probability theory, and descriptive Bayesianism should be thought of as the claim that the former are rare (even if the latter are rampant).

Psychologists have mustered a welter of evidence showing that people do indeed tend to obey Bayesian norms, at least in certain tasks. Still the most impressive piece of evidence to date stems from Mike Oaksford and Nick Chater's (1994) work on the previously mentioned Wason selection task. Whereas Wason and other early psychologists of reasoning had thought of this and similar tasks as *deductive* inference tasks, Oaksford and Chater convincingly argued that they are better conceived as *probabilistic* inference tasks. Specifically with respect to the selection task, these authors showed that when thus conceived, the predominant response Wason and colleagues had found was in fact the one we should expect to find if people followed something close to an optimal Bayesian strategy for acquiring information. The core idea is that people will tend to interpret the task before them as one of discovering a statistical dependence between A cards and 2 cards, and that by turning the A card and the 2 card in front of them, they are gaining the most informative evidence regarding such a dependence; at least, that is so given assumptions about people's priors and about how best to measure information gain, which Oaksford and Chater argue to be plausible in the given task.[10] Other frequently cited evidence apparently supporting descriptive Bayesianism is found in later work by the same authors (e.g., Oaksford & Chater, 1996, 2007; Chater & Oaksford, 1999) and in the work of Joshua Tenenbaum and colleagues (e.g., Griffiths & Tenenbaum, 2006; Tenenbaum, Griffiths, & Kemp, 2006; Gopnik & Tenenbaum, 2007; Tenenbaum et al., 2011).

10. It is to be noted, though, that results reported in Nickerson, Butler, and Barch (in press) suggest that lack of motivation to engage with the materials and, more importantly still, a tendency to misinterpret "If there is an A on one side of the card, then there is a 7 on the other side" as "If there is an A on the visible side of the card, then there is a 7 on the hidden side" (and similarly for similar statements) may also be part of the explanation of participants' performance in typical selection-task experiments.

However, these apparent successes merit two comments. First, although there is the previously presented evidence showing that people at least in certain tasks seem to follow Bayesian prescriptions, there are also well-documented violations of those same prescriptions. Among these is the so-called conjunction fallacy (Kahneman, Slovic, & Tversky, 1982; Tversky & Kahneman, 1983), that is, the finding that people sometimes deem a conjunction more probable than its least probable conjunct—which is a violation of item 3 of proposition 3.1 above, and an interesting one at that, inasmuch as we generally have no difficulty recognizing that a conjunction entails its conjuncts. However, it has been said (e.g., by Tversky and Kahneman themselves) that this is a case in which people confuse a question about probability with one about representativeness, or in which they mistake probability for the related but different concept of confirmation (Tentori, Crupi, & Russo, 2013). That would not make the fallacy any less fallacious—given that probability is neither representativeness nor confirmation—but at least the fallacy would no longer directly contradict descriptive Bayesianism. Similarly, researchers have recorded order-effects on sequential probability judgments, effects that, normatively speaking, should not occur (see Baratgin & Politzer, 2007, and further discussion in this chapter). Here too, attempts have been made to explain these effects in ways that render them compatible with descriptive Bayesianism. I further mention the so-called central tendency effect, which seems to show a general preference for middling values, including middling degrees of belief (i.e., degrees of belief not too far from the midpoint of the scale on which they are usually measured), which again can lead people to hold degrees of belief that are not formally probabilities (Stevens, 1971). Again, there may be a Bayesian explanation for this effect (Douven, 2018). Perhaps most problematic is the base rate fallacy (Bar-Hillel, 1980), which is famously illustrated by the fact that people experiencing certain symptoms can panic if an online search informs them that some dreadful disease is typically accompanied by those symptoms, even if the prior probability for anyone contracting that disease is vanishingly small. It shows that people have a tendency to equate Pr(Disease | Symptoms) with Pr(Symptoms | Disease), in clear violation of Bayes's theorem, as stated in the previous section.

The second comment is that it is not always easy to distinguish experimentally between descriptive Bayesianism and what one might more broadly call "descriptive probabilism," according to which degrees of belief are central to our reasoning, where these degrees of belief also tend to be approximately

representable as probabilities and where updates of our degrees of belief tend to approximately proceed by Bayes's rule (Over, 2009; Elqayam & Evans, 2013; Elqayam, 2018). To understand why that matters, recall from the preceding chapter that instances of EXPL could be said to be close to Bayes's rule, certainly for small values of the explanation bonus that is attributed. In any event, much of the evidence supposedly supporting descriptive Bayesianism does not and was not intended to discriminate between descriptive Bayesianism and descriptive probabilism. What is more, Oaksford and Chater (2013, p. 374) explicitly admit that "it is unclear what are the rational probabilistic constraints on dynamic inference," where by "dynamic inference" they mean updates of the kind that according to strict Bayesian norms are to be governed by Bayes's rule and related rules, such as Jeffrey conditionalization.

Evidence that people update their degrees of belief in ways inconsistent with Bayes's rule dates back at least to the 1960s and hence long predates the advent of the New Paradigm. Lawrence Phillips and Ward Edwards (1966) used a new at the time experimental paradigm to show that people's probability estimates after the receipt of new evidence differed systematically from what they should have been according to Bayes's rule. In this paradigm, participants are informed that one container (e.g., an urn or a bag) holds two different types of objects (e.g., red chips / blue chips, or black balls / white balls) in one specific ratio, and another container holds those same types of objects in a different ratio. The participants are then shown a collection of objects sampled randomly from one of the containers, without being told from which, and are asked to estimate the probability that the sample comes from the first container rather than from the second. That the estimates deviate reliably from Bayesian updates has been replicated in numerous experiments since (see, e.g., Edwards, 1968; Marks & Clarkson, 1972; Fischhoff & Lichtenstein, 1978; Schum & Martin, 1982).

These deviations from Bayes's rule could have been, and perhaps were, entirely unsystematic. In any event, none of the early experimental work addressing the descriptive adequacy of Bayesian updating probed the possibility that people's updates were influenced by explanatory considerations. Needless to say, that possibility is of immediate interest to the topic of this book, and we next look at work concerned with the possible connection between belief updates and explanation.

3.3 Explanation and Belief Change

The first psychologists to seriously look into the connection between updating and explanation were Nancy Pennington and Reid Hastie in their influential work on juror decision making (1988, 1992, 1993). This work showed the importance of the order in which evidence is presented to jurors. Pennington and Hastie's participants were reliably more inclined to judge a defendant guilty when the prosecutors presented their evidence in an order that facilitated the mental construction of an explanatory story of how the crime unfolded. These researchers also found evidence that how the different pieces of evidence impacted their participants' degrees of belief differed consistently from what descriptive Bayesianism predicted.

Also closely related to a comparison between Bayesian and explanatory reasoning, Bénédicte 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. Although they were not thereby violating Bayes's rule (which is not meant to cover the provision of statistical information, as previously mentioned), they *were* violating the Principal Principle (see previous discussion), which many Bayesians endorse.

According to Bes et al., the effect of the causal information on their participants' probability judgments is likely related to the extent to which the participants could process that information into an explanatory story. It was not part of these authors' design to ask participants to judge the explanation quality of the statements whose probabilities they were asked to estimate. Such judgments could have been illuminating and might well have revealed a strong correlation between their participants' explanation-quality judgments and their probability judgments.

That, at least, is suggested by the experimental results from work that Jonah Schupbach and I documented in Douven and Schupbach (2015a). In that paper, we compared the descriptive adequacy of Bayesianism with the descriptive adequacy of explanationism (see p. 20), that is, the claim that, in updating their degrees of belief, people take into account their explanatory judgments in a way not captured by Bayes's rule. In particular, we were interested in the following questions:

(1) How do Bayesianism and explanationism compare with regard to their descriptive adequacy? Do judgments of the explanatory goodness of hypotheses play an essential role in updating, in a way that is incompatible with the Bayesian doctrine?
(2) If explanatory judgments are found to have such a role, do conditional probabilities retain an important influence in updating alongside such judgments?
(3) What sort of explanatory judgments in particular (if any) factor into updating?

To answer these questions, Schupbach and I re-analyzed data from an experiment first reported in Schupbach (2011), which used the "bookbags-and-poker-chips" paradigm from Phillips and Edwards's earliest studies on probabilistic updating.

However, Schupbach's experiment added a twist to that paradigm. Schupbach designed a sequential probabilistic updating task that allowed him to measure the degree to which explanatory factors influenced the updating. He interviewed the participants individually, showing them at the start two urns, urn A and urn B, each of which contained forty balls. The participants were also shown that urn A contained thirty black balls and ten white ones and that urn B contained fifteen black balls and twenty-five white ones. This information stayed available to the participants, in the form of a picture of the urns' contents, during the whole interview. Then on the basis of the outcome of a coin flip, one or the other urn was chosen, outside of the participants' view, after which ten balls were drawn, one by one, from the selected urn and were lined up before the participants. The balls were drawn without replacement, which was also clearly visible to the participants. The participants were asked to answer three questions after each draw. In order, these were

(i) how well the hypothesis that urn A had been selected explained the results from the drawings so far;
(ii) how well the hypothesis that urn B had been selected explained those results; and
(iii) how probable it was, in their opinion, that urn A had been selected, in view of the results so far.

Participants had to answer the questions about explanatory goodness by marking a point on a continuous scale with anchors "extremely poor explanation" and "extremely good explanation."

To analyze the responses, Douven and Schupbach (2015a) used mixed-effects models, a type of statistical model briefly mentioned in chapter 1. Such models are becoming increasingly popular in various areas of research, including psychology (see, e.g., Baayen, 2008, ch. 7; Gelman & Hill, 2009; Zuur et al., 2009). As discussed on page 11, mixed-effects models allow you to focus on the variance in the data in which you are interested, and filter out variance in your data that results from random elements in your design. For instance, the group of participants that you have is typically a random matter. Had you run your experiment a day earlier or a day later, or had you offered to pay your participants less or more, then presumably you would have had an at least partly different group of participants, which in all likelihood would have resulted in an at least partly different data set. Specifically with respect to the research reported in Douven and Schupbach (2015a), it is reasonable to think that judgments of explanatory goodness may relate to posteriors differently for different participants. Mixed-effects models allow one to take such possible individual differences into account by treating participants as so-called random effects.

In the main part of our analysis, Schupbach and I fitted a number of mixed-effects models, each with the collected participants' responses to the third question as fixed effect and at least the objective conditional probabilities as predictor variable. The models also had random intercepts and slopes for participants. In a classical regression model there is exactly one intercept—which predicts the outcome variable when all predictor variables are 0 (or, if the variables have been scaled, at their mean)—and one slope, which predicts how the outcome variable will change depending on changes in the predictor variables. By contrast, in a mixed-effects model each random effect (participants, in Douven and Schupbach's models) can have its own intercept and slope.

Three of the mixed-effects models reported in Douven and Schupbach (2015a) are of particular interest to us. In the first of these (model MMO, in Douven & Schupbach, 2015a), 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 (MMOAB). The third model of interest had next to objective conditional probabilities the computed *difference* between the participants' responses to the first question and their responses to the second question as predictor (MMOD). Because including more predictors will in general lead

Table 3.1: Model comparison results concerning the main models from Douven and Schupbach (2015a).

	k	LL	AIC	ΔAIC	BIC	ΔBIC
MMO	6	282.34	−552.68	82.37	−531.32	68.12
MMOAB	15	329.13	−628.26	6.79	−574.85	24.59
MMOD	10	327.53	−635.05	0.00	−599.44	0.00

Note: *k* is the number of parameters and LL the log-likelihood. ΔAIC is the difference in AIC value with the model with lowest AIC value, and analogously for ΔBIC.

to better model fit, these models were compared using criteria that penalize for extra predictors, in particular the so-called Akaike information criterion (AIC) and the Bayesian information criterion (BIC), which weigh (in slightly different ways) model fit, as measured in terms of the log-likelihood of the model (basically, how much variance remains unaccounted for by the model), against model complexity, as measured in terms of the number of parameters estimated in the model (such as the coefficients of the fixed effects in the model and the variance of the residuals; see Akaike, 1973, and Schwarz, 1978; for useful discussion, also see Burnham & Anderson, 2002, and Sober, 2002).[11] Taking model complexity into account helps one reduce the risk of *overfitting*, that is, fitting one's model so tightly to the *available* data that it generalizes poorly to *new* data (where generalization to new data is typically what one is interested in). The results from the model comparisons are summarized in table 3.1.

AIC values and BIC values are to be interpreted as penalties, meaning that lower is better. Also, they make sense only comparatively, and then only for

11. It could be said that thereby AIC and BIC are a bit in the spirit of abduction. But note that the idea of abduction is broader, in that it is not confined to statistical model selection. More importantly, there is no reason for explanationists to commit to an unpacking of simplicity in terms of number of adjustable parameters or to the view that simplicity is the only or dominant factor determining explanation quality. Finally, whereas AIC and BIC are often presented as based on an analytical result, it is important to notice that this result depends on a number of substantive assumptions. In particular, it is assumed that the distance of a fitted model from the truth is to be measured by the Kullback–Leibler distance. As shown in chapter 7, there are alternatives to that, and a key finding of that chapter suggests that which distance to use may depend on the problem at hand. Accordingly, while to the best of my knowledge this has not been investigated, there is a real possibility that, depending on context, different information criteria, which weigh fit and complexity differently, are appropriate.

models that are fit to the same data, as is the case here. The AIC value of MMO was more than 10 higher than those of the other models, which according to Kenneth Burnham and David Anderson (2002) is to be interpreted as indicating that the former model enjoys *no* support from the data, given the availability of the other models. The pattern is the same for the BIC values.[12]

These model comparisons would seem alarming for advocates of descriptive Bayesianism, who hold that Bayesian norms also achieve greater predictive accuracy than competing norms. After all, again by the Principal Principle, Schupbach's participants after each draw should have set their degree of belief that urn A had been selected equal to whatever the objective conditional probability was that that urn had been selected, given all registered draws at that point. It is not just that this outcome failed to materialize. The bigger problem is that whereas from a Bayesian perspective it should appear puzzling that the participants' judgments of explanatory goodness reliably helped to predict their degrees of belief, by explanationists precisely this was predicted. And from the perspective of anyone committed to the broad idea of abduction, there also *should* have been an impact of explanatory considerations on people's updates.

Thus, the answer to the first research question from Douven and Schupbach (2015a) is that explanationism is predictively more accurate than Bayesianism and that judgments of explanatory goodness factor in people's updates in an essentially non-Bayesian way. At the same time, it is not as though conditional probabilities played no role. To the contrary, as Schupbach and I reported, conditional probability came out as highly significant in every model that contained it as a predictor, answering our second research question in

12. Given that the argument is between explanationists and Bayesians, it is worth rerunning the analysis using the machinery of Bayesian statistics. In Bayesian statistics, the currently recommended criterion for model comparison is the so-called leave-one-out cross-validation information criterion (LOOIC), which is, as with AIC and BIC, to be thought of as a penalty: a lower value indicates greater predictive accuracy. The Bayesian analysis showed that the equivalent of model MMO has an LOOIC value of –607.3 and the equivalent of model MMOD has an LOOIC value of –691.7. A Bayesian analysis also allows one to calculate weights for each of the models, which can be interpreted as the probability that the given model will perform best on new data, conditional on the supposition that the true model is among the ones considered (Vehtari, Gelman, & Gabry, 2017). Performing the calculations for those models showed that, in light of Douven and Schupbach's data, the model with objective conditional probabilities and difference in explanatory goodness as predictors has a weight of almost 90 percent, whereas the "Bayesian" model, with only objective conditional probabilities as predictor, has a weight of just over 10 percent.

the positive. The answer to the final question is especially noteworthy. In MMOD, the mixed-effects model with predictors (1) objective conditional probabilities, and (2) difference in explanatory goodness between the hypothesis that urn A had been selected and the hypothesis that urn B had been selected, both predictors came out highly significant. Moreover, this was the best model. So, the answer to the third question is that, to the extent that explanatory considerations influence updating, the salient considerations will be comparative, conveying the extent to which an explanatory hypothesis outperforms its competitor(s). This is noteworthy because it is exactly what one would expect to find supposing the descriptive adequacy of Bird's suggestion, mentioned in section 2.2.2, that in abductive reasoning we should heed not only the quality of the best explanation but also how much better it is, qua explanation, than the second-best explanation. We encounter more direct support for that suggestion subsequently in this chapter.

Whereas the preceding corroborates the hypothesis that explanatory considerations are at play when people update, at least in Schupbach's experimental setting, it is consistent with everything said so far that *how* explanatory considerations influence updating is unsystematic at least to the extent that it cannot easily be conceived as the result of rule-following behavior, for instance, of following something like EXPL. However, results from Douven and Schupbach (2015b) suggest that people respond to the receipt of new evidence at least somewhat as if they followed an update rule akin to EXPL.

In that paper, Schupbach and I used the objective probabilities from our 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 instead urn B had been selected). We then tried to predict the participants' updates again, now using objective conditional probabilities and computed explanatory goodness of the two hypotheses as predictors. So, we basically repeated the analysis summarized previously, but now using, instead of the *subjective* judgments of explanatory goodness that had served as predictors in the earlier analysis, the *computed* explanatory goodness values.

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, Karl 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 I. J. Good's (1960) measure, according to which H's power to explain E equals[13]

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

In other words, if only objective probabilities are available to help us predict people's updates, we can substantially improve on the Bayesian model by making the objective probabilities do double duty and use them also as input in Popper's or Good's measure of explanatory goodness. At least this is so in the kind of updating context studied in Douven and Schupbach (2015a).

The results from Douven and Schupbach (2015b) serve further in the book as inspiration for delving deeper into the normative aspects of the debate between Bayesians and explanationists.

3.4 Just Noise?

We showed that there have been attempts to explain away previously known violations of Bayesian norms. But together these form a rather heterogeneous set of proposals. It is only very recently that authors have tried to explain those violations in a unified manner, attributing all or at least a great number of them to a cognitive mechanism we know to be active on independent grounds. Specifically, Fintan Costello and Paul Watts (2016, 2018) have sought to explain away deviations from Bayesian norms in terms of "noisy" sampling from memory. The deviations addressed by these authors were not directly related to the *updating* of degrees of belief, but there is nothing in their approach per se to suggest that it might not apply equally to registered deviations from Bayes's rule, such as those reported in the previous section. What would thereby be achieved is *not* that somehow what appear to be deviations fall into place as being in accordance with Bayesian updating after all. Instead, although there would still be evidence of a failure of descriptive Bayesianism, it would be of an utterly boring kind, not *really* brought about by some influence of

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

explanatory considerations on reasoning but simply due to the well-recorded fact that human memory is fallible. In other words, the violations would still be real but of a kind that we should expect to observe even if people did what they could to meet Bayesian standards. If so, then this would show that descriptive Bayesianism can still be correct—to the extent that any theory of rationality can be descriptively correct, given some long-known human imperfections.

At the root of Costello and Watts's work is the eminently plausible contention that people make random errors in estimating probabilities, including estimating conditional probabilities. For example, on their account we would estimate the conditional probability that a student will pass a given exam on the supposition that he or she studies hard by sampling from memory students who worked hard for an exam and then counting among those the ones who passed the exam. But this process is error-prone: our memories are not fully dependable, and we may thus misremember a student who worked hard but failed to have passed the exam, or the other way around; indeed, we may even misremember a student to have worked hard for an exam. Costello and Watts (2018, p. 9) make the general assumption that "events have some chance $d < 0.5$ of randomly being read incorrectly." They demonstrate that for a number of probabilistic identities, such errors tend to cancel out, whereas for others they do not or even get compounded. They offer an abundance of evidence for their hypothesis and show how that helps to explain a number of well-known biases, including order effects in sequential probability judgments and the conjunction fallacy.

The aim of Costello and Watts is not to show that, contrary to what the mainstream believes, these and related biases are unproblematic. Rather, it is to identify their source, namely, the noise present in the process by which we determine probabilities. Thus, the claim is, the said biases do not refute descriptive Bayesianism, given that this position is not committed to the assumption that human memory is failsafe. Because "our reasoning processes are necessarily subject to noise" (Costello & Watts, 2016, p. 131), any alternative to descriptive Bayesianism faces the same problem that it can be accurate only up to the biasing effects of sampling errors.

As said, Costello and Watts do not consider deviations from Bayes's rule.[14] But they are right to remark that their results present a challenge that goes beyond non-Bayesian explanations of the biases that they did consider:

> [R]andom noise in reasoning can cause systematic biases in people's responses even when people are using the rational reasoning processes of standard frequentist probability estimation. To demonstrate conclusively that people are using heuristics, researchers must show that observed biases cannot be explained as the result of systematic effects caused by random noise. (Costello & Watts, 2016, p. 132)

Surely the same is true for attempts to explain biases by reference to mechanisms other than heuristics, such as attempts to attribute the findings summarized in the previous section to the influence of explanatory considerations on people's updating. Might that ostensible influence not also be in reality attributable to estimation errors?

First, there are some a priori reasons to doubt that Douven and Schupbach's results are due to the kind of sampling noise that figures in Costello and Watts' account. As said, that noise is supposed to arise from the fact that we estimate probabilities, including conditional probabilities, by sampling and counting from memory, and in that process sometimes make mistakes. Although plausible in general, it is to be recalled that in the experiment reported in Schupbach (2011), participants at all times had the drawings from the urns with their contents in front of them and were allowed to consult this memory aid as often as they wanted. At a minimum, their estimation of conditional probabilities would seem to have carried a lesser risk of being affected by noise than if these conditional probabilities had been the result of sampling from memory.

Second, adopting Costello and Watts's proposal to account for the deviations from Bayes's rule recorded by Douven and Schupbach would raise an explanatory challenge. For then where would the predictive superiority of the explanatory models come from? Given that in Costello and Watts's model the noise is supposed to be *random*, how could it moderate people's probability judgments precisely in such a way that in Douven and Schupbach's

14. That is, they do not consider deviations from what *we* are calling "Bayes's rule." They do consider what they call "Bayes rule identities," by which they mean the most direct consequences of the definition of conditional probability, including (what we call) Bayes's theorem.

analysis their explanatory judgments would come out as highly significant predictors? There does not appear to be anything in Costello and Watts's model that could account for a close link between the (putative) random noise at work in people's probability estimates and those same people's judgments of explanatory goodness.

Raising a priori doubts about the applicability of Costello and Watts's proposal to the data at issue takes us only so far. After all, while it is uncontroversial that human memory is error prone, researchers have identified various sources of noise in the nervous system that can have behavioral consequences, including causing information processing slips (Faisal, Selen, & Wolpert, 2008). So, random noise may corrupt the estimation of conditional probabilities even if that estimation does not involve any sampling from memory. And for the Bayesian purpose of explaining away the seeming influence of explanatory considerations, the exact provenance or nature of the noise is immaterial. Indeed, while Costello and Watts's assumption that the errors they measured are due to memory glitches may be plausible, there is nothing in their papers to exclude that those errors are not actually, wholly or partly, of a different origin.[15]

Thus, a better way to answer the question of whether Costello and Watts's model can account for Douven and Schupbach's findings is to try to predict the subjective updated probabilities reported by the latter authors, not by the objective conditional probabilities (henceforth called O) but instead by the "noisy" version of that predictor, transformed according to a formula given by Costello and Watts.

To do so, we first define a function that takes Costello and Watts's error parameter d as input and outputs the transform of the predictor O, to be designated as $f(\mathrm{O}, d)$, for the given value of d. The function f is based on equation 17 in Costello and Watts (2018), according to which

$$\Pr_*(\varphi \mid \psi) \;=\; \frac{(1-2d)^2 \Pr(\varphi \wedge \psi) + d(1-2d)\big(\Pr(\varphi) + \Pr(\psi)\big) + d^2}{(1-2d)\Pr(\psi) + d},$$

where $\Pr_*(\varphi \mid \psi)$ is the noisy estimate of the probability of φ conditional on ψ, and $d \in [0, .5)$ is the noise parameter. Notice that if $d = 0$, indicating that there is no noise at all, then $\Pr_*(\varphi \mid \psi) = \Pr(\varphi \wedge \psi)/\Pr(\psi) = \Pr(\varphi \mid \psi)$.

15. Nor would it matter to their overall conclusion. See the previous citation in this section, in which they refer to random noise in reasoning in general.

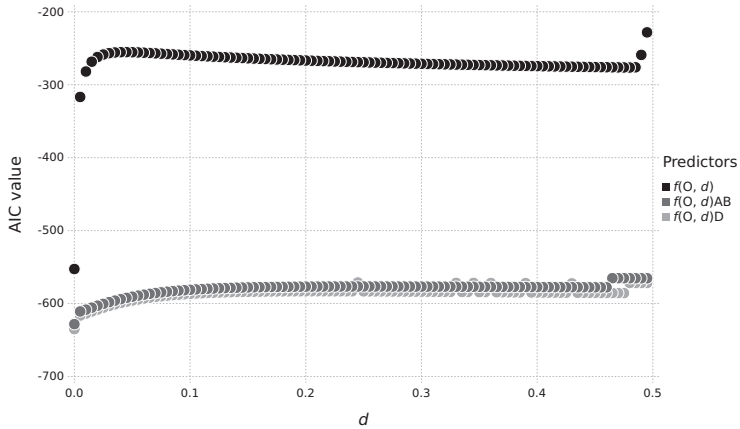

Figure 3.1: AIC values of the models with the same predictors as MMO, MMOAB, and MMOD, respectively, except that O is replaced by $f(O, d)$, for d from 0 to 0.5 in steps of 0.005.

We can then fit, for d from 0 to 0.5 in increments of 0.005, mixed-effects models like the ones specified in the previous section except that the new models have the noisy objective conditional probabilities $f(O, d)$ rather than the untransformed ones, O, as a predictor. If the deviations from Bayes's rule reported in Douven and Schupbach (2015a) are due to the interference of noise in registering frequencies, in the manner of Costello and Watts, then we should find that for some values of d the corresponding Bayesian model with $f(O, d)$ as its only predictor does best, and adding judgments of explanatory goodness as predictors should not lead to any improvement.

Figure 3.1 shows the AIC values of all the fitted models. It is immediately obvious that replacing O by $f(O, d)$, for any admissible value of d, only leads to worse fit and that in particular for no value of d does the Bayesian model—or at least the model that does not take judgments of explanatory goodness into account—come close, in terms of model quality as captured by AIC, to the models that do take such judgments into account (whether directly or indirectly via the difference in goodness between the two urn hypotheses). Comparison in terms of BIC values leads to qualitatively identical results. (See the Jupyter notebook for this chapter that is part of the Supplementary Materials; see appendix E for instructions on use.)

In short, not only is there a priori little reason to believe that Costello and Watts's proposal might also hold the key to understanding the deviations from Bayesian updating to be found in the data from Douven and Schupbach (2015a), but there is also no support from the data.

3.5 Explanatory Reasoning and Accuracy

So far, we have looked at a re-analysis of the data from Schupbach's (2011) experiment. Those data were gathered in one-on-one interviews, and so the number of participants was understandably limited (twenty-six, to be precise). That was at least mildly concerning in itself, but furthermore, because each participant had received a unique series of draws of balls from the selected urn, it gave no reliable information about individual differences, notably, about how widely the influence of explanatory factors on updating varied across participants.

To that end, Douven and Schupbach (2015a) conducted a follow-up experiment. This experiment was conducted online and had 259 participants. The participants were divided into four groups, and each group received a different one of four series of draws that had been randomly chosen from the twenty-six series that had occurred in Schupbach's experiment. Just as in that experiment, participants were informed that the balls would be drawn without replacement from an urn that had been chosen from two urns—again called "urn A" and "urn B"—on the basis of the flip of a fair coin. They were also fully informed about the contents of each urn. They were then shown, one at a time, the ten draws and asked, after each draw, to make the judgments 1, 2, and 3 as previously described (on p. 84).

Of Douven and Schupbach's 259 participants, 206 remained after exclusion of participants on the basis of a number of standard criteria (such as being nonnative speakers of English), and 167 remained after further participants were excluded for failing to answer some comprehension questions correctly. We regressed the remaining participants' degrees of belief that urn A had been selected, first onto the objective probabilities that urn A had been selected, and second onto those objective probabilities as well as the difference between (1) the judged explanatory goodness of the hypothesis that urn A had been selected and (2) the judged explanatory goodness of the hypothesis that urn B had been selected. This led to results that were qualitatively identical to those for the main experiment: the larger model vastly outperformed the "Bayesian"

model with only objective probabilities as independent variable, and it did so across all conventional measures of model fit. Most notably, the Bayesian model had an AIC value of –1627.55 and the larger, "explanationist" model had an AIC value of –2422.07. (Recall that for AIC values, lower is better.)

At several junctures it was already indicated that Bayesians tend to greatly emphasize the importance of making updates minimally inaccurate. That sounds reasonable enough: we want to have the sharpest picture of reality possible, and our updating practices should help us obtain it. We go into the details of the Bayesian inaccuracy-minimization argument in chapter 4 and will find it wanting. For now, all we need to know are two things. The first is that Bayesians are really concerned with *expected* inaccuracy (i.e., the inaccuracy that, from your current perspective, you expect to achieve by updating), having left the question of *actual* inaccuracy entirely to the side.[16] The second is that the most popular measure of inaccuracy (according to some Bayesians, the only correct one) is the so-called Brier score. We formally define and discuss this scoring rule in chapter 5 and here mention only that Brier scores are penalties and that hence lower Brier scores are better.

Douven and Schupbach were interested in evaluating the descriptive adequacy of Bayesianism and explanationism; they were not concerned with the issue of accuracy. However, in Douven (2016b) I brought their data to bear on that issue, as follows. For each of the four series of draws used in Douven and Schupbach's follow-up experiment, it was known from which urn draws had come in the experiment from Schupbach (2011), in which people were interviewed individually. Specifically, three of the groups had received series of draws that had come from urn B, and the remaining group had received a series of draws from urn A. Consequently, for each of the participants in the follow-up experiment I could calculate the Brier score incurred over the ten draws. I could then run a separate regression analysis for each participant, again with degrees of belief after each update as dependent variable and objective probabilities and difference in judged explanatory goodness between the hypotheses as independent variables. The β weights (basically, standardized regression coefficients) for the independent variables thereby obtained can be interpreted as the relative weights a participant assigns to probability and explanation in determining their new degree of belief after seeing the outcome of a draw. Finally, I checked how these weights related to how accurate, in

16. Briggs and Pettigrew (2020) is an exception, but see chapter 4, footnote 17.

terms of Brier penalties, the given participant was. The question I thus sought to answer was whether taking into account explanatory considerations enables people to achieve greater accuracy, or whether it has an opposite effect, or whether there is no relation at all.

To determine the relationship between, on one hand, the weights that the participants gave to probability and to explanation and, on the other, the total Brier score that they incurred, I fitted a number of additional linear models, which had as dependent variable the total Brier scores of each participant and as independent variables one or both of these participants' β weights for probability and the β weights for explanation that resulted from the individual regression analyses mentioned previously. It turned out that the model with both the β weights for probability and those for explanation as independent variables topped the others in terms of all model comparison criteria. In that model, the slope for the probability variable was −1.91 and the slope for the explanation variable −2.38 (the regression results were all highly significant; see Douven, 2016b, for statistical details). This means that both variables had a lowering (i.e., improving) effect on Brier scores. Specifically, the best model indicates that, keeping the weight for probability fixed, every extra unit of weight given to explanation (as measured in standard deviations) lowered the total Brier score by over two points on average. The effect of attending to objective probabilities in the model was although similarly directed somewhat smaller. So, supposing that objective probabilities are available, it is certainly a good idea to attend to them, in line with what normative Bayesians would recommend. However, the more crucial finding of the analysis from Douven (2016b) was that contrary to what those Bayesians would recommend, it is an even better idea to attend at the same time to explanatory considerations, as doing so is likely to further increase the accuracy of one's degrees of belief.

3.6 Good-Enough and Second-Best Explanations

In the foregoing, we have concentrated on experimental results bearing on probabilistic versions of abduction, such as EXPL. Such versions also take the limelight in subsequent chapters. It is nevertheless worth briefly discussing experimental work in which Patricia Mirabile and I addressed a number of questions more directly having to do with categorical versions of abduction (Douven & Mirabile, 2018). We were especially (although not exclusively) interested in the descriptive adequacy of ABD2, proposed by Musgrave and Lip-

ton, which emphasized the need for the best explanation to be good enough, and of Bird's proposed amendment of that version, which adds that the best explanation should be considerably better than its closest competitor. Specifically, we focused on the following research questions:

(1) Is the quality of an explanation a good predictor of people's willingness to accept that explanation, and is there a threshold effect in that an explanation's quality must be above a certain threshold for people to accept it?
(2) Given a potential explanation of some phenomenon, does it make any difference, in regard to the perceived quality of that explanation and in regard to people's willingness to accept it, whether or not a second explanation is introduced (and if so, why)?
(3) Given two rival explanations of some phenomenon, does the *magnitude* of their difference in quality make a difference to people's willingness to accept the better of the two?

We conducted three experiments that together were designed to answer these questions.

All three experiments drew on a fixed set of six basic scenarios, each of which presented a fact accompanied by one (in the first experiment) or two (in the second and third experiments) explanations of that fact. These explanations could vary in quality; where two explanations appeared, the explanations could vary in quality independently of each other. Participants were asked whether they accepted one of the explanations, how likely they thought the explanations were, and how good they were, qua explanations. In the third experiment, they were also asked, for each scenario separately, how confident they were in their judgment about the acceptability of the explanation.

Across three experiments, we found evidence that the quality of an explanation is a good predictor of people's willingness to accept that explanation.[17] There was also clear evidence of a threshold effect, supporting the descriptive adequacy of Lipton's proposal. Thus, the answer to the first of the preceding questions is *yes*. Interestingly, the first two experiments also showed that

17. Basically for the reasons mentioned in footnote 12 in this chapter, we ran both frequentist and Bayesian analyses for all experiments. The results of the frequentist and the Bayesian analysis were always in line with each other. See Douven and Mirabile (2018) for details.

although an explanation's probability was a good predictor of willingness to agree with the explanation, it was as such inferior to explanation quality.

It was further found that it matters for people's willingness to infer the truth of an explanation whether that explanation appears alone or side by side with a competitor, although it does *not* affect how they perceive the quality of the explanation, thus answering the first part of research question 2 in the negative and the second part in the positive.[18] The same results showed that the quality of the competitor matters to people's willingness to infer the truth of the other explanation: if an explanation was presented together with a competitor that was about as good, participants tended to be significantly less inclined to infer the truth of the former, whereas the effect of introducing a competitor tended to be smaller when that competitor was clearly inferior. That is a *yes* to the third of the research questions.

The data from Experiment 3 in Douven and Mirabile (2018), specifically participants' answers to the questions about their confidence in their own acceptability judgments, offered a good insight into exactly *why* the presence and quality of a second candidate explanation matter to people's willingness to infer to the best explanation. It could have been that those factors had somehow influenced people's *perception* of the quality of the alternative explanation. Contrast effects have long been known in perception research—a color can look brighter in the presence of a second, less bright color than the first color looks on its own, to mention a famous example—but more recently these effects have been shown to occur also in the cognitive domain and in particular to affect people's standards of judgment. For instance, it has been found that whether a consumer product is judged to be desirable may depend on which other products a potential buyer is attending to (Shoots-Reinhard et al., 2014). The dampening effect that the presentation of a second explanation had could have been due to something similar: confronted with a second possible explanation, people might no longer have found the first explanation

18. Tenney, Cleary, and Spellman's (2009) work on the notion of being beyond reasonable doubt in the context of criminal law already gave some reason to expect these results. In an experiment, these authors found that the introduction of an additional suspect by the defense in a (fictional) murder case significantly lowered their participants' preparedness to find the target suspect guilty beyond reasonable doubt. While Tenney and colleagues do not relate their results to explanatory reasoning, it is natural to think of the introduction of the second suspect as offering an alternative explanation for the events described in the case.

so compelling, which could explain why they were less inclined in that case to infer to that first explanation.

As previously mentioned, however, it was found that the perception of the quality of the explanations was largely unaffected by the presence of an alternative alongside them. Instead, the alternative lowered people's inclination to infer to the other explanation, at least when the alternative was about equally good or not much worse, because that undermined people's confidence in the inference.[19] The said dampening effect thus appeared to be akin to a seemingly unrelated result reported by Ruth Horry and Neil Brewer (2016). These authors found that when participants were briefly shown a face and then asked to identify that face in a set of faces shown simultaneously, the confidence in their choice correlated negatively with the degree of similarity between the target face and whichever other face in the set was most similar to it.

It should probably have been unsurprising that we found something similar in our experiment: given two explanations competing for bestness, the more similar in quality they are, the less confident people become in their judgment of which is best and the more hesitant they become to infer the truth of that explanation. After all, it was noted earlier that explanation quality is supposed to depend on such factors as simplicity, scope, and coherence with background knowledge. And each of those factors carries some vagueness with it. So, explanation quality being somewhat vague, it is easily understandable that people may feel that their verdicts concerning the relative quality of explanations are not entirely reliable whenever they are confronted with explanations similar in quality. This explains an empirical finding, but at the same time it provides a rationale for Bird's proposal: explanatory bestness being somewhat vague, a judgment to the effect that this or that explanation tops the others will not always be reliable, especially not if we are confronted with two or more explanations that are close in terms of quality. Then it would seem reasonable to refrain from inferring to the best explanation, or at a minimum from having much confidence in the conclusion.

19. As, following Ackerman and Thompson (2015, 2017, 2018; also Thompson, Prowse Turner, & Pennycook, 2011), we could more exactly say: it undermined their *metacognitive* confidence, that is, their confidence in (parts of) their own cognitive functioning, in this case their ability to draw the right conclusion based on explanatory considerations.

3.7 Should Philosophers Care?

This chapter has reported several empirical studies related to abduction—for the most part, probabilistic versions thereof. These studies have all been published in psychology journals and fit into the more general trend in those journals of an increasing number of papers devoted to explanation. And there is certainly more empirical work to be done. As Douven and Schupbach (2015a) remark, it is a limitation of their study that the hypotheses and contexts that they examined were quite simple and artificially stochastic in nature and that their materials did not, for instance, provide any causal-mechanical details describing how the explanandum came about. That means that the explanatory judgments that their participants gave have inevitably been of a rather shallow sort. They thus call for follow-up research using less artificial, explanatorily richer contexts. Douven and Mirabile (2018) used precisely such contexts, but they looked at categorical versions of abduction only and also did not study sequential updating.

Is there any reason for *philosophers* to be interested in the work reported in the foregoing, or in what follow-up research may bring? No doubt some will shrug their shoulders, declaring that because there is no inference from "is" to "ought," there is no reason for philosophers, who are not empirical scientists and should be concerned only with normative matters, to even look at any of the empirical data relating to abduction. Yet I believe that many philosophers *will* be interested, although their reasons may vary.

First, I mentioned the growing enthusiasm for empirically informed philosophy, which seems to be largely fueled by a concern that analytic philosophy is at the risk of devolving into the study of Alice and Bob. You know, Alice and Bob, who are constantly wondering whether they really saw Carl or rather his evil twin Dave—or was it their friend Eve in disguise? (Eve loves practical jokes; just the other day, she dressed as a painted mule.) And on it goes, until Zachary enters the picture, at which point we are all supposed to have the feeling that we are onto something deep. Thinking up such little stories can be tremendous fun and can sometimes also lead to real insights, but it can also be unsatisfactory in the long run, if only because we often just *know* that soon someone will come up with a different story, possibly starring Alfie and Beatrice, that undermines whatever lessons we thought could be learned from Alice and Bob. Thus people began to look for more solid starting points in philosophy, out of which grew empirically informed philosophy, its aim being

to replace the intuitions that Alice-and-Bob stories were supposed to elicit with hard data about what laypeople, not indoctrinated by years in graduate school, thought was reasonable to say under various circumstances.

Engaging in empirically informed philosophy usually meant one of two things: sifting through the psychology journals, hunting for results that could be related to this or that philosophical dispute; or getting one's own hands dirty, by conducting experiments, typically ones that asked for ordinary people's verdicts about cases that philosophers had strong opinions about (like Gettier cases or the kind of cases figuring in the skepticism debate). This approach has led to quite a bit of bad news for philosophers lately, in that there appeared to be surprisingly little support among nonphilosophers, for claims that philosophers had variably designated as intuitive, pretheoretically obvious, natural to say, or using similar expressions meant to signal that a claim does not stand in need of any serious argumentative backing (e.g., Machery et al., 2004; Weinberg et al., 2010).

Not all philosophers agree that these findings should be taken seriously (see, e.g., Williamson, 2007, 2018). However that may be, it is certainly fortunate to have some *good* news for a change, in that with regard to abduction, the empirical results are in line with ideas that philosophers had put forward as being intuitively correct. This is true for Bird's as well as Lipton's proposals, both of which concerned a *categorical* version of abduction, but also for the idea that explanatory considerations are central to how we update our *graded* beliefs. Naturally, however, this will still give limited reason to rejoice if you deem, as some do, the said philosophical proposals completely wrongheaded.

This brings me to another reason why I expect at least some philosophers to be interested in the data discussed in this chapter. Even if there is no inference from the descriptive to the normative, it is still true that few philosophers relish attributing massive error to the folk; indeed, it has been argued that we should do so only if all else fails (Wilson, 1959; see also the quote from Lewis, 1986, on p. 58). Still, attributing massive error is what we must do, given what we examined in this chapter; at least, it is what we must do if the mainstream is right and abduction is to be repudiated as irrational. There may be considerable evidence that people sometimes behave like Bayesian agents, but there is also considerable evidence that sometimes they do not and that at least sometimes when they do not, the reason is that their belief changes are guided by perceived explanatory goodness. This puts some pressure on the

mainstream to rethink its stance vis-à-vis abduction and revisit its arguments against abduction.

To be sure, there may be no way to escape those arguments. But in the following I argue that that is just not so. As a matter of fact, I want to show that the way in which those arguments fail is anything but subtle or otherwise hard to grasp—which may raise the question of why they are so widely taken for granted. On my analysis, these arguments use a kind of diversionary tactic not unlike the old magician's technique that makes the audience look *here* so that it will not pay attention to the real action (dumping a pack of cards, smothering a rabbit, stitching together a female assistent, . . .) that is going on *there*; specifically, these arguments make us focus so strongly on potential *costs* of explanatory reasoning that they make us forget to ask about potential *benefits*. This is not to suggest that the diversion was surreptitiously *planned* by the authors who came up with these arguments. It may simply be an unintended side effect of the arguments' fairly technical nature, which may have created a kind of collective blindspot.[20] Be this as it may, that we should finally start attending to abductive reasoning's advantages is a recurring message in the next chapters.

20. See the literature on attentional blindness (e.g., Chabris & Simons, 2010) for an explanation of how that may come about.