6

The Ecological Rationality of Abduction

Google wasn't that much faster than Yahoo, but it didn't need to be. All it needed to be was a little bit faster and the rest is history.

—Stanley Druckenmiller

6.1 The Justification of Abduction

To this point the book has been concerned with staving off the main criticisms leveled against abduction. Nothing presented so far amounts to a *justification* of abduction. Although, as we have discussed, there is strong evidence that in some contexts people do update their beliefs by abduction, most of what was presented in the foregoing chapters is compatible with the thought that the effect of explanation on updating is no more than the manifestation of a bias. From here onward, I aim to convince you that updating one's beliefs, or degrees of belief, on the basis of explanatory considerations can be the right thing to do. Specifically, the claim argued is that explanation-based updating allows people to strike the best balance between being fast and being accurate learners.

By doing so, I expand on work begun in chapter 4, which compared Bayes's rule and EXPL along the dimensions of speed and accuracy. There, the comparison served only to rebut the claim that any form of non-Bayesian updating is to be rejected for failing to minimize inaccuracy. No conclusion was drawn regarding whether Bayes's rule or EXPL would be the more rational update rule to use. And what conclusion *could* have been drawn? The results from the earlier comparison appeared to be mixed. EXPL was found to be

better at picking up the signal in the data, but when noise occasionally took a pattern that made it look like a signal, it was the first to be misled, sometimes quite dramatically.

If the question whether Bayes's rule or EXPL is the rational pick appears hard to answer, then, I submit, that is only because we are committed to an outdated "universalist" conception of rationality, which implies among other things that there must be exactly one right way to update one's beliefs or degrees of belief. Instead, I aim to follow recent work in psychology and cognitive science that argues for a conception of *ecological* rationality, an arguably more realistic and useful conception of rationality that leaves open the possibility that, at different times and for different people, different practices may count as rational. Authors advocating this conception have laid out clearly how one should go about investigating the rationality of cognitive practices. Following their guidance, I use computer simulations implementing an agent-based optimization technique to compare Bayes's rule with EXPL and similar probabilistic forms of abduction. The results will show that in some stylized but still realistic contexts, abduction, and not Bayes's rule, is the rational choice of update rule. But before embarking on a new defense of abduction, I should briefly explain why I believe that previous defenses fall short.

6.2 Previous Defenses of Abduction

For reasons already noted (see the last paragraph of chapter 2), not many today would want to subscribe to a conception of truth that posits a necessary connection between explanatory force and truth—for instance, because it stipulates explanatory superiority to be necessary for truth, as do some pragmatist conceptions of truth. At a minimum, that makes the prospects of an a priori defense of abduction dim. And indeed, all better-known defenses of abduction are of an empirical nature in that they appeal to data suppos-

edly supporting the claim that (in some form) abduction is a reliable rule of inference.[1,2]

A particularly well-crafted defense of this sort was developed by Boyd in the 1980s (see Boyd, 1984, 1985, 1990). It starts by underlining the theory-dependence of scientific methodology, which comprises methods for designing experiments, for assessing data, for choosing between rival hypotheses, and so on. For instance, in considering possible confounding factors from which an experimental setup has to be shielded, scientists draw heavily on already accepted theories. These inform them about which confounds to reckon with and which steps can be taken to prevent those confounds or to control for them. The argument next calls attention to the apparent reliability of this methodology, which after all has yielded and continues to yield impressively accurate theories. In particular, by relying on this methodology, scientists have for some time been able to find ever more instrumentally adequate theories. Boyd then argues that the reliability of scientific methodology is best explained by assuming that the theories on which it relies are at least approximately true. From this and from the fact that these theories were arrived at mainly by abductive reasoning, he concludes that abduction must be a reliable rule of inference.

Boyd's is a *brave* attempt at defending abduction. It does not engage in the kind of appeasement politics of those who seek to show that abduction can be regarded as serving a convenient helper function for Bayesians, as discussed in section 1.2.2, but rather aims to show that abduction is a mode of inference in its own right and as such forms the core of scientific methodology. But critics have accused Boyd's argument of being circular (e.g., Laudan, 1981; Fine, 1984). Specifically, it has been said that the argument rests on a premise—that scientific methodology is informed by approximately true background theories—that in turn rests on an abductive step for its plausibility. And the reliability of this type of inference is precisely what is at stake. At a more

1. Biggs and Wilson (2017) do think abduction can be justified a priori. But their argument involves the claim that conditionals of the sort "If water and H_2O are spatiotemporally coincident, then water is identical with H_2O" can be justified a priori by means of what they call "a hypothetical form of abduction" (p. 750). I fail to see, however, what support there is for the reliability of that form of abduction.

2. Even though Devitt (1991, p. 111) does not think abduction is a priori, he still finds it so obviously correct that he thinks no defense is called for. Naturally, this was written before Bayesian epistemology became dominant and even came to appear "obviously correct" to many.

general level, Jonathan Vogel (2008, p. 531) has argued that the belief that an epistemic rule *R* is reliable cannot be justified by the application of *R*.

To this, Psillos (1999, ch. 4) has responded by invoking a distinction credited to Richard Braithwaite, to wit, between *premise circularity* and *rule circularity*. An argument is premise circular if its conclusion is among its premises. A rule-circular argument, by contrast, has a conclusion that asserts something about an inferential rule that is used in the same argument. In Psillos's analysis, Boyd's argument is rule circular but not premise circular, and rule-circular arguments, Psillos contends, need not be viciously circular (even though a premise-circular argument is always viciously circular). To be more precise, in his view, an argument for the reliability of a given rule *R* that essentially relies on *R* as an inferential principle is not vicious, provided that the use of *R* does not guarantee a positive conclusion about *R*'s reliability. Psillos claims that in Boyd's argument, this proviso is met. Although Boyd concludes that the background theories on which scientific methodology relies are approximately true on the basis of an abductive step, the use of abduction itself does not trivialize his conclusion. After all, granting the use of abduction does nothing to ensure that the best explanation of the success of scientific methodology is the approximate truth of the relevant background theories. Thus Boyd's argument still stands, or so Psillos maintains.

Even if the use of abduction in Boyd's argument might have led to the conclusion that abduction is *not* reliable, the argument's rule circularity might still be worrisome. Suppose that some scientific community relied not on abduction but on a rule that we may dub "Inference to the Worst Explanation" (IWE), a rule that sanctions inferring to the *worst* explanation of the available data. We may safely assume that the use of this rule would lead mostly to the adoption of very unsuccessful theories. Nevertheless, the said community might justify its use of IWE by dint of the following reasoning: "Scientific theories tend to be hugely unsuccessful. These theories were arrived at by application of IWE. That IWE is a reliable rule of inference—that is, it usually leads from true premises to true conclusions—is surely the worst explanation of the fact that our theories are so unsuccessful. Hence, by application of IWE, we may conclude that IWE is a reliable rule of inference." Whereas this would be an utterly absurd conclusion, the argument leading up to it cannot be convicted of being viciously circular any more than Boyd's argument for the reliability of abduction can (if Psillos is right). It would appear then that there must be something else amiss with rule circularity.

It is fair to note that, for Psillos, the fact that a rule-circular argument does not guarantee a positive conclusion about the rule at issue is not sufficient for such an argument to be valid. A further necessary condition is "that one should not have reason to doubt the reliability of the rule—that there is nothing currently available which can make one distrust the rule" (Psillos, 1999, p. 85). And there is plenty of reason to doubt the reliability of IWE; in fact, the previous argument *supposes* that it is unreliable. Two questions arise, however. First, why should we accept the additional condition? Second, do we really have *no* reason to doubt the reliability of abduction? Certainly *some* of the abductive inferences we make lead us to accept *falsehoods*. How many falsehoods may we accept on the basis of abduction before we can legitimately begin to distrust this mode of reasoning, at least enough to fail Psillos's extra condition? No clear answers have been given to these questions.

Be this as it may, even if rule circularity is neither vicious nor otherwise problematic, one may still wonder how Boyd's argument is to convert a critic of abduction, given that it relies on abduction. But Psillos makes it clear that the point of philosophical argumentation is not always to convince an opponent of one's position. Sometimes the point is, more modestly, to assure or reassure oneself that the position that one endorses, or is inclined to endorse, is correct. In the case at hand, we need not think of Boyd's argument as an attempt to convince the opponent of abduction of its reliability. Rather, it may be thought of as justifying the rule from within the perspective of someone who is already sympathetic toward abduction. This is an important insight to which I come back in chapter 8.

There have also been attempts to argue for abduction in a more straightforward fashion, to wit, via enumerative induction (e.g., Harré, 1986, 1988; Bird, 1998; Kitcher, 2001; Douven, 2002a). The common idea of these attempts is that every newly recorded successful application of abduction—like the discovery of Neptune, whose existence had been postulated on explanatory grounds, as discussed in section 2.1—adds further support to the hypothesis that abduction is a reliable rule of inference, in the way in which every newly observed black raven adds some support to the hypothesis that all ravens are black. Because it does not involve abductive reasoning, this type of argument is more likely to also appeal to disbelievers in abduction.

One problem with the argument is precisely that it relies on induction, whose status has been questioned as much as that of abduction has. Did David Hume (1748/2006) not show that a justification of induction is unattainable?

According to him, a *deductive* justification is not forthcoming because induction is not *necessarily* reliable, whereas an *inductive* justification is ruled out on grounds of circularity.

However, abduction was unknown to Hume, at least as an explicitly recognized mode of inference, so he could hardly have considered the idea of induction and abduction taking in each other's wash. Prima facie, it certainly makes sense to think that we might be able to base a defense of induction on abduction *and* vice versa. It would appear, after all, that we have reason to be confident in the reliability of induction because the registered successes of induction are *best explained* by its reliability, while conversely we have reason to be confident in the reliability of abduction because the registered successes of abduction provide strong *inductive evidence* that abduction is reliable. There is no apparent rule circularity here, nor any other kind of circularity.

Alas, this defense would fare no better than Psillos's defense of abduction. Consider the following duo of inferential principles: IWE as previously introduced and Wesley Salmon's (1963) rule of counterinduction, according to which from the observation that of all Φs observed so far x percent were Ψs it follows that of all Φs yet to be observed $100 - x$ percent will be Ψs. Although both principles are patently nonsensical, we could argue for them in tandem, as follows. We can ask why counterinduction fails, or would fail, so dismally, and then surely the worst answer we could give is that it is a reliable inferential principle. But then IWE licenses the inference that the rule of counterinduction is reliable indeed. That is helpful only if we can justify IWE. And we can, namely, by invoking counterinduction. After all, we can be sure that IWE would, if actually in use, have let us down time and again. By counterinduction, that means that it would do just great in the future!

The kind of construction used here, in which one hypothesis (the reliability of IWE) serves as an auxiliary in helping to confirm a second hypothesis (the reliability of counterinduction) while the second hypothesis serves as an auxiliary in helping to confirm the first, has received some attention from confirmation theorists under the heading "bootstrap confirmation" (e.g., Glymour, 1980; Christensen, 1983, 1997; van Fraassen, 1983; Douven & Meijs, 2006; Douven & Kelp, 2013). The proposal of bootstrap confirmation came with an important additional condition, however: given a theory, thought of as a set of hypotheses, it should not only be the case that each hypothesis is confirmed by the evidence supposing one or more of the other hypotheses

as auxiliaries—or even more debatably (Douven & Meijs, 2006), supposing itself as an auxiliary—but it should *also* hold that the fact that parts of the theory are allowed to help other parts in obtaining support from the evidence should not lead to *trivialization*, meaning that it should not shield the theory from disconfirmation. We can gain support for a theory only if we put that theory at risk of being disconfirmed.

In the case at hand, however, the nontrivialization condition is not going to help, for the joint assumption of IWE and counterinduction does *not* prevent the occurrence of disconfirming evidence. Our reliance on counterinduction in this argument does not by itself prevent disconfirmation of IWE, given that our judgments of explanation quality might be so badly off that by going for the worst explanation we often end up endorsing what is by more reasonable standards the best explanation, in which case IWE might have yielded good results, and an application of counterinduction would have led us to dismiss IWE. Similarly, we may assume IWE and yet live in a hopelessly counterinductive universe, in which case counterinduction might have worked and the worst explanation of that would have been that counterinduction is unreliable, so that IWE would have led to a disconfirmation of counterinduction.

But even if it is difficult to make the deficiency of the bootstrap defense of the combination of abduction and induction formally precise, that such a defense *is* deficient is still clear from the fact that, by parallel reasoning, we could argue for a pair of what we said are patently nonsensical inferential principles.

Recent progress on the justification of induction raises hope that a defense of abduction may proceed via induction, but that a defense of induction then does not need to appeal to abduction in turn but can be accomplished by recruiting *meta-induction*. In a series of papers (Schurz, 2008b, 2009, 2012, 2017), and most comprehensively in his 2019 book, Schurz argues that Hume, and all authors who have since studied the question of the justification of induction, have erred in assuming that such a justification must involve a demonstration of the *reliability* of induction.[3] Instead of focusing on its reliability, we should focus on the question of its *optimality*. Suppose in all possible circumstances we could only be worse off by *not* relying on induction.

3. See also Schurz and Thorn (2016) and Thorn and Schurz (2019); see Sterkenburg (2019, 2020), Douven (in press), and Schurz (in press) for discussion.

Then obviously we would be justified in relying on induction. In fact, doing so would appear rationally mandated in that case.

Taking his cue from work on prediction with expert advice (in particular, Cesa-Bianchi & Lugosi, 2006), Schurz first shows that meta-induction, understood as induction over object-inductive methods, does have an analytic justification, consisting of a proof that, in all possible circumstances, the meta-inductive strategy that aggregates the predictions of the available object-inductive methods according to their registered successes is at least as accurate as and under weak additional assumptions even more accurate than any of those object-inductive methods. Whereas this gives us a justification only of meta-induction, the justification of object-induction now follows meta-inductively from the fact that object-inductive methods are known to have served us better than any non-inductive methods have. It is important to be clear about how this proposal differs from the naïve inductive justification of induction, which was already considered by Hume and rejected by him and most authors since. Although the naïve justification also starts from the registered successes of our inductive methods, it then applies *object*-induction to *object*-inductive methods. That presupposes an antecedent warrant for applying object-induction, but whether we have that warrant is precisely the question at issue. By contrast, Schurz applies *meta*-induction, which is *provably* optimal, to *object*-inductive methods. No circularity is involved in this argument.

Schurz's work on induction marks a major step forward, and I do think that it can be instrumental in defending abduction as a rational mode of inference. Granting this, there is still a concern about the *scope* of an inductive justification of abduction. Such a justification presupposes that we can agree on what counts as a successful application of the rule. In the case discussed in section 2.1 of the head and headless body that were found close to each other in Jackson, Mississippi, the medical examiner's report confirming what many had already inferred could be regarded as an undisputed piece of evidence for the reliability of abduction. But abduction is also used to infer the existence of unobservable entities—as in Johnson's "discovery" of the electron—and unobservable processes and mechanisms, as in Darwin's "discovery" of the mechanism of variation and selective retention. I put scare quotes around "discovery" in the previous sentence precisely because scientific antirealists will deny that we are in a position to claim that the electron has been discovered, or that the mechanism of variation and selective retention has been discovered.

More generally, scientific antirealists will want to remain agnostic regarding existence claims for any unobservable entities or processes. It would seem that, as a result, we can perhaps inductively justify applications of abduction as long as those applications concern strictly observable matters, but that we could not go beyond that without facing the charge of question-begging. For surely antirealists will argue that we are already relying on abduction if we cite any evidence presupposing the existence of unobservables as supporting the reliability of abduction also when it pertains to hypotheses postulating the existence of unobservables. Worse yet, by much the same reasoning, the Cartesian skeptic could challenge even the alleged evidence in the head-in-Jackson case, given that it already presupposes the existence of an external world, which the skeptic will refuse to grant.

This is a *serious* concern, insofar as the prime philosophical use of abduction has been in the context of underdetermination claims, discussed in section 2.3. And both the skepticism debate and the scientific realism debate were said to turn on such claims. The scientific antirealist is willing to grant a greater epistemic access than the skeptic, but both hold that it is more limited than people ordinarily think: the data we do have access to (sensory impressions in the case of the skeptic and data about observables in the case of the scientific antirealist) do not allow us to draw any conclusions about an external world (according to the skeptic) or about the unobservable part of the world (according to the scientific antirealist), precisely because, to their eyes, any such conclusion would be underdetermined by the evidence—that is, by what, from their perspective, can be legitimately designated as such. That is where opponents have always wanted to wheel in abduction: perhaps there are rival hypotheses that are equally compatible with the evidence, but the external world hypothesis, and scientific realism, offer better *explanations* of the evidence. Even if that is so, however, the skeptic and the antirealist will remain unfazed as long as they see no grounds for believing abduction to be a reliable rule of inference, and the question is how we are to show, without begging any questions against either the skeptic or the antirealist, that it is.

We will return to this question in chapter 8. Right now we are still facing a more fundamental challenge. For the broadly Bayesian arguments considered in chapter 4 charge abduction not just with having limited applicability but also with being inapplicable *tout court*, on pain of irrationality. We saw that the arguments failed: they proceeded by claiming that abductive reasoning could be costly but then forgot to ask whether it might be able to deliver

goods that are worth the costs (granting the costs are real to begin with). At the same time, we have only scratched the surface when it comes to whether we can indeed get benefits from abduction that we cannot get from Bayes's rule nor perhaps from any other rule, and more broadly, whether it could ever be rational to reason abductively. I now want to turn to these questions.

It is important, however, to be clear about the notion of rationality that I assume in the following. Thinking about rationality was long dominated, and in philosophy still *is* dominated, by the assumption that norms of rationality should be *universal*: they should apply to everyone alike in every imaginable situation. The debate then was about what those universal norms were. As explained in section 3.2, the once popular view that the norms of rationality are the rules of logic was replaced by the psychologically more plausible Bayesian view that sees rationality as essentially about managing uncertainties. Although this was a step forward, various authors have argued that the Bayesian ideal of rationality, as formalized by the principles stated in section 3.2.1, is still woefully unrealistic. In particular, it has been argued that this ideal is unachievable for limited beings like us (see, e.g., Giere, 1988, ch. 6; Arkes, Gigerenzer, & Hertwig, 2016) and that it is not even clear what steps could be taken to *approach* it (Earman, 1992, p. 56). Such concerns have led various researchers to doubt the validity of Bayesianism as a basis for human rationality and to pursue alternatives more closely in touch with evolution-oriented theories of cognition.

This alternative approach came to fruition most prominently in the development of Simon's (1982) celebrated theory of bounded rationality and Gerd Gigerenzer and collaborators' related theory of ecological rationality (e.g., Gigerenzer & Goldstein, 1996; Gigerenzer et al., 1999; Gigerenzer, 2000; Goldstein & Gigerenzer, 2002; Gigerenzer, Hertwig, & Pachur, 2011; Todd & Gigerenzer, 2012; Todd & Brighton, 2016; Schurz & Hertwig, 2019). A more recent proposal in this vein is Elqayam's (2011, 2012) account of grounded rationality.[4]

Although these accounts differ in their details, they have the important commonality of taking into consideration the various biological and cognitive limitations to which humans are subject, as well as the environment or environments in which we, both individually and socially, operate and pursue

4. For a related account, called "resource rationality," see Griffiths, Lieder, and Goodman (2015).

our personal interests. The said limitations impose constraints on rationality that are to some extent universal—we are *all* finite beings, with finite computational powers, finite memory capacity, and so on—but the environment in which we are to act is not the same for everyone and may be different for each of us at different points in time; the same holds for the cognitive tools available to us and for the goals that we pursue. Gigerenzer and colleagues and also Elqayam argue that, on their accounts, there can be no universally applicable rationality standards that we might be able to pin down a priori. Whether a person's behavior qualifies as rational depends on whether the behavior facilitates achievement of the person's particular goal or goals in the context at issue and given the resources available to that person in that context.

As Elqayam has pointed out, the term "behavior" in the present conception of rationality is to be understood broadly, so as to include *cognitive* behavior. Concomitantly, the person's goals may be partly or wholly cognitive ones. She thus sees her notion of grounded rationality as in accordance with Jonathan Evans and David Over's (1996) proposal to conceive of instrumental (pragmatic, goal-oriented) rationality as primary and as subsuming epistemic rationality, given that among a person's goals may be obedience to certain systems of epistemic norms (like classical logic or the laws of probability).[5] But grounded rationality goes beyond a mere allegiance to instrumental rationality in that its emphasis on the role of context and on individual differences strongly implies that "instrumental rationality cannot be reduced to any one-size-fits-all normative framework" (Elqayam, 2012, p. 46; also Elqayam & Evans, 2011).

Adopting this more psychologically oriented approach to rationality makes it even clearer that as argued in chapter 4, the dynamic Dutch book argument against non-Bayesian update rules cannot simply be dismissed on the grounds that it deals with the wrong notion of rationality: epistemic rationality is a *species* of pragmatic rationality. It is then equally clear, however, that a verdict about the rationality of a given update rule cannot be reached without considering possible users of such rules *as located in a given context, with their specific goals in that context and their specific cognitive capacities*. Only then will it make sense to judge the use of a particular update rule as

5. That epistemic grounds are a special type of instrumental grounds has also recently been argued in Steglich-Petersen and Skipper (2019) and, from the standpoint of social epistemology, in Dyke (in press).

rational, or as more rational than the use of some alternative rule, and the judgment will have to be based on a cost–benefit analysis of the rule or rules within that context. To put this another way: on the present conception of rationality, the question of whether a given update rule is rational or not cannot be answered in the abstract. Given a person and an environment, we can ask which rule is most likely to help the person succeed; and given an update rule, we can ask when and for whom it is most likely to be helpful.

Hal Arkes, Gerd Gigerenzer, and Ralph Hertwig (2016, p. 33) are very specific about how one should investigate the rationality of a cognitive strategy. Their proposal is to begin by identifying (1) the goal of an individual or a group; (2) the strategies available to the individual or group for achieving that goal; and (3) the structural properties of the individual's or group's environment. In a second step, one then determines which strategy or strategies are most likely to help the individual or group achieve her / its goal in the given environment. Note how different this procedure is from trying to lay down rationality criteria a priori.

6.3 Simulating Explanatory Reasoning

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

6.3.1 Setup

In chapter 2, I stated EXPL as a schematic probabilistic explication of abduction, which gives some positive bonus to best explanations, though I have used the label mostly to designate what is actually only a specific instance of the schema with that label, the instance in which the explanation bonus c was 0.1. From now on, I use the label again to refer generically to instances of the schema with $c \in (0, 1)$ and so to instances that can be said to embody some probabilistic form of explanatory reasoning. To these we add the instances of two further schemata, inspired by the finding from Douven and Schupbach

(2015b) that Good's and Popper's measures of explanatory goodness (stated in section 3.3) were useful in predicting participants' responses from one of the studies reported in Douven and Schupbach (2015a).

Any instance of EXPL gives all credit for explanatory goodness to the hypothesis that explains the evidence best and gives nothing to any other hypothesis. The aforementioned objective measures of explanatory goodness can help to give content to the claim that probabilistic versions of abduction may instead credit each hypothesis separately, in proportion to that hypothesis's explanatory power. EXPL was seen to work very much like Bayes's rule except that it adds a bonus point to the best explanation and then normalizes to obtain probabilities again. We cannot quite follow this procedure if we want to credit all hypotheses separately, at least not if the crediting is to be done on the basis of Good's or Popper's measure. In particular, given evidence E and a set $\{H_i\}_{i\leqslant n}$ of mutually exclusive and jointly exhaustive hypotheses, we cannot first update the hypotheses on E via Bayes's rule, then add $\mathcal{M}(H_i, E)$ to H_i (with $\mathcal{M}$ a measure of explanatory goodness), and finally normalize. The measures of objective goodness considered have a range of $[-1, 1]$ (see ch. 3, footnote 13) with 0 being the neutral point. This means that actually they can assign bonus points as well as malus points; where a hypothesis is an extremely poor explanation of the evidence, they can even assign a malus point of -1, which when added to the hypothesis's probability could result in a negative value unsuitable for "normalizing" to a probability.

Therefore, the rules to be considered will first update the hypothesis's probability via Bayes's rule, then calculate the hypothesis's explanatory goodness according to a given objective measure, next add or subtract a percentage of the hypothesis's probability in proportion to its explanatory goodness, and finally renormalize. More formally, the rules that we consider are instances of the following schema:

$$(6.1)\quad \Pr'(H_i) = \frac{\Pr(H_i)\Pr(E|H_i)+c\Pr(H_i)\Pr(E|H_i)\mathcal{M}(H_i,E)}{\sum_{j=1}^{n}\left(\Pr(H_j)\Pr(E|H_j)+c\Pr(H_j)\Pr(E|H_j)\mathcal{M}(H_j,E)\right)},$$

where Pr and Pr′ are as defined previously. What I call "Good's rule" is obtained from this by substituting Good's measure for $\mathcal{M}$, and what I call "Popper's rule," by substituting Popper's measure for the same expression. Note that, as stated, these rules are also mere schemata from which specific instances are derived by fixing a particular value for $c \in (0, 1)$, which determines what percentage of H_i's probability after a Bayesian update on E is

added in proportion to this hypothesis's power to explain E. Here, too, we would obtain Bayes's rule by setting $c = 0$.[6]

The rules compared in the following—the available strategies, in the terminology of Arkes, Gigerenzer, and Hertwig (2016)—are Bayes's rule and instances of EXPL, Good's rule, and Popper's rule, where for all these instances $c \in (0, 0.25)$. The setting in which these rules are compared is, although stylized, not at all unrealistic. It concerns medical doctors at an intensive care unit (ICU) who try to determine what exactly is wrong with the patients who are rushed in, and who, on the basis of the information that they obtain from the tests administered, must choose an intervention. They are under some pressure to act fast: as time passes, the probability that the patient will die increases, and making the right intervention *decreases* that probability. On the other hand, making the wrong intervention *increases* probability of death.

Within this broad setting, we look at two more specific environments. In the first—the "Weibull environment"—we assume that for all patients brought in, probability of death can be roughly modeled by the cumulative density function (CDF) of some Weibull distribution. Such distributions are characterized by two parameters, a shape parameter k and a scale parameter l, and the associated CDF is given by

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

More exactly, we assume that the probability of death for a given patient can be modeled in terms of an instance of this schema, where for each patient, k and l are chosen randomly, with $k \sim \mathcal{U}(0.5, 5)$ and $l \sim \mathcal{U}(50, 250)$. The resulting distribution is commonly designated as "Weibull(k, l)."

In the second environment—the "Gamma environment"—we assume instead that probability of death can be modeled by the CDF of some Gamma distribution. These distributions are also characterized by a shape parameter k

6. In terms of Vassend's (in press) distinction between inferential and predictive updating, Good's and Popper's rules are inferential update rules whereas EXPL is a predictive update rule. In the same paper, Vassend distinguishes between legitimate and nonlegitimate update rules, with Good's and Popper's rules falling into the former category and EXPL into the latter. It is to be noted, however, that he makes commutativity—in the sense that the order in which we evaluate different pieces of evidence should make no difference to the overall result—a necessary condition for legitimacy, and that is a debatable assumption (Lange, 2000).

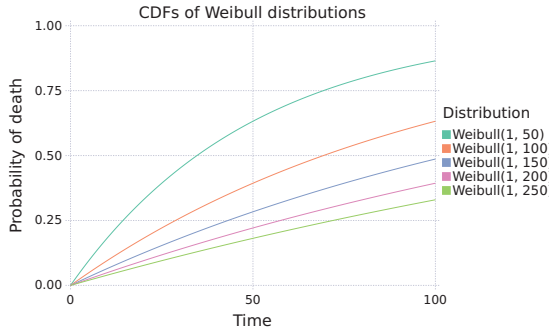

Figure 6.1: Examples of Weibull CDFs that give the probability of death of a patient as a function of time after admission into the intensive care unit.

and a scale parameter l, and the associated CDF is

$$F(x; k, l) \quad = \quad \frac{1}{\Gamma(k)} \gamma\left(k, \frac{x}{l}\right),$$

where the function Γ is a generalization of the factorial function and γ is the so-called lower incomplete gamma function.[7] The resulting distribution is commonly designated as "$\Gamma(k, l)$." Figures 6.1 and 6.2 show some examples of CDFs of Weibull and Gamma distributions, respectively.

Also associated with each patient are two parameters $a \geqslant 1$ and $b \geqslant 1$ that indicate the effect on probability of death of the right intervention and the effect on that probability of a wrong intervention, respectively. Where p_t is the probability of death of a given patient at time t, performing the right intervention at t lowers the probability of death to p_t/a whereas performing a wrong intervention at t raises that probability to $(b + p_t - 1)/b$. Figure 6.3 illustrates these effects for the case where $a = b = 2$, both for a specific Weibull distribution (left) and for a specific Gamma distribution (right).

We further assume that what is wrong with the patient can be expressed in terms of one parameter, α; we imagine that at the time the patient is rushed into the ICU her relevant medical status is known up to the value of this parameter. At that time, the only thing known about α is that it can take a value in $\{0, .1, .2, \ldots, 1\}$, with none of these values initially more likely

7. Specifically, $\gamma(a, x)$ is defined to be $\int_0^x t^{a-1} e^{-t} \mathrm{d}t$.

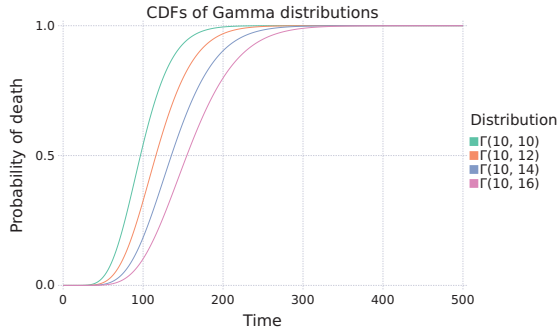

Figure 6.2: Examples of Gamma CDFs that give the probability of death of a patient as a function of time after admission into the intensive care unit.

than any other. To estimate the value of α, the doctor has to rely on the test results that she receives, with one new result coming in per unit of time. The results are either "positive" or "negative," and the tests are probabilistically independent of each other and all have the same (unknown) probability of

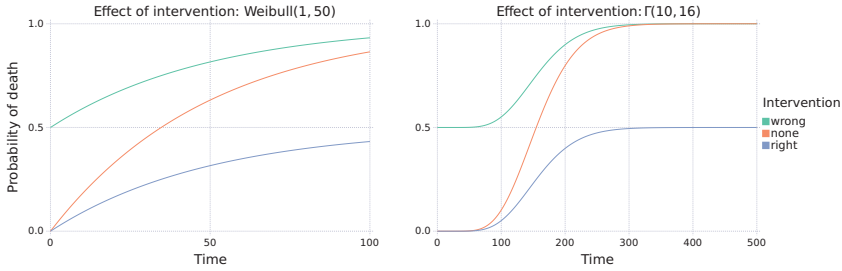

Figure 6.3: Examples of the effect of right and wrong interventions for a Weibull distribution (*left*) and for a Gamma distribution (*right*). In both panels, the orange graph is the probability of death of the patient over time if no intervention is performed; the green graph gives, for every point in time, the probability of death of the patient if at that point in time a wrong intervention is performed; the blue graph does the same for the correct intervention.

being positive. The hypothesis that $\alpha = x$ is to be interpreted as implying that the probability for any given test turning up positive equals x.

Stephen Pauker and Jerome Kassirer (1980) and Benjamin Djulbegovic and colleagues (2014) found that the so-called threshold model, according to which physicians should decide to administer treatment precisely if the probability of disease is above a specified threshold, accurately predicts decision making in clinical practice (see also Djulbegovic & Elqayam, 2017). In line with this finding, we assume that the doctor must be at least 90 percent certain about a hypothesis before she is prepared to intervene. It is further stipulated that only if the doctor comes to believe the true hypothesis to a degree above that threshold will she perform the correct intervention; coming to believe a false hypothesis to a degree above the threshold will lead to an incorrect intervention, with any incorrect intervention having the same detrimental effect (as specified by the parameter b) on the probability that the patient will survive.

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

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

6.3.2 Method

We can think of this question as a constrained optimization problem.[8] Even though all relevant features of the setting have been characterized mathematically, this problem appears too complex for analytical methods to give much guidance. However, we can resort to an optimization technique known as "agent-based optimization," which is a form of genetic programming (Holland, 1975; Koza, 1992; Bäck, 1996; Yu, Yao, & Zhou, 2012; Kochenderfer & Wheeler, 2019, ch. 9). This technique takes its inspiration from the principles of natural selection and genetics, letting agents represent different solutions to a given problem, determining their "fitness" level (according to some criterion of fitness deriving from whatever problem needs to be solved), and then selecting, either deterministically or stochastically, the fittest agents, which are retained and / or allowed to reproduce in some predetermined way and thereby provide the input population for the next round, in which the competition for survival or reproduction starts again. This can be repeated over and over, possibly as often as is needed to obtain a fixed point at which all agents represent the same solution (see, e.g., Barbati, Bruno, & Genovese, 2012; also various papers in Czarnowski, Jędrzejowicz, & Kacprzyk, 2013). Agent-based optimization has been applied with notable success to optimization problems in a broad variety of fields, including chemistry, economics, medicine, operations research, psychology, and robotics. (See, for instance, Sarkar & Modak, 2005; Dhanalakshmi et al., 2011; Heris & Khaloozadeh, 2011; Douven, 2019b; and, for a general overview, Coello Coello, Lamont, & Van Veldhuizen, 2007, ch. 7.) As subsequently shown, using this technique has the additional advantage of allowing us to shed some light on how evolution may have contributed to shaping our inferential practices and in particular may have led to the adoption of certain forms of explanatory reasoning.

I have conducted the same agent-based optimization procedure for each of the two environments introduced previously. I first give an informal description of this procedure. The procedure started with a population consisting of 200 "medical doctors," with fifty of them using Bayes's rule, fifty using

8. *Constrained* optimization because we are not looking for the optimal update rule per se (optimal in the context) but the rule among those that we are *considering* that best fits the context. This makes the current approach different from Anderson's (1990, 1991) rational analysis, which starts with basically the three steps from Arkes and colleagues summarized at the end of the previous section and then adds as an additional step a requirement to look for the optimal procedure *tout court* for achieving the individual's or group's goal.

an instance of EXPL, fifty using an instance of Good's rule, and fifty using an instance of Popper's rule; for the non-Bayesian updaters, the value of the explanation bonus c was chosen randomly per doctor, with $c \sim \mathcal{U}(0, 0.25)$. Then each doctor was assessed on the basis of treating one hundred patients, where the relevant characteristics of the patients (survival probability, effects of right and wrong interventions, and value of α) were chosen randomly and separately per patient in the way specified above (so, for instance, in the Weibull environment, the survival probability was based on a randomly chosen Weibull distribution, with the parameters falling within the indicated bounds).

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

After each doctor had treated her one hundred patients, her total score was calculated, and the 50 percent fittest doctors (the doctors with the highest average patient survival rate) were determined. These one hundred doctors were retained for the next generation, and they were also allowed to replicate by having a copy of themselves in that generation. In every simulation, the previous steps were repeated for 250 generations, and in total fifty simulations were run.

To make this formally precise, we think of doctors as ordered triples $\langle r, b, \Pr^{r,b} \rangle$, where r is one of the rules defined above; b is a bonus value, which is either 0, for Bayes's rule, or some random value in $(0, 0.25)$ for the other rules; and $\Pr^{r,b} = \langle \Pr_0^{r,b}, \ldots, \Pr_{100}^{r,b} \rangle$ is the sequence of the doctor's

degrees-of-belief functions, starting with the degrees-of-belief function $\Pr_0^{r,b}$ that the doctor has when the patient is brought in, before receiving any test results, and with $\Pr_t^{r,b}$ being the doctor's degrees-of-belief function at time $t \in \{1, 2, \ldots, 100\}$, when she has received, and sequentially updated on, $\mathcal{E}_t = \langle E_1, \ldots, E_t \rangle$, with E_t the test result received at time t, and where $\Pr_t^{r,b}$ comes from $\Pr_{t-1}^{r,b}$ by updating on E_t via rule r with bonus value b (conventionally set to 0 in case the rule is Bayes's rule).

Patients can be thought of as ordered quintuples $\langle \alpha, k, l, a, b \rangle$ with all parameter values chosen randomly (within certain bounds) per patient. The first parameter, α, concerns what is wrong with the patient (think, for instance, of the value of some blood parameter), and it is unknown at the time when the patient is brought in; the only thing known at that time is that $\alpha = h$ for some $h \in \{0, 0.1, \ldots, 1\}$, with each of those possibilities equally likely at that time; thus, $\Pr_0^{r,b}(\alpha = h) = 1/11$, for all h, r, b. The true hypothesis concerning α gives rise to the evidence (the test results) that the doctor receives during the time of the treatment, in the precise sense that $\alpha = h$ if and only if the long-run frequency of positive outcomes equals $h \times 100$ percent. The k and l parameters are the shape and scale parameters of the Weibull distribution in the Weibull environment and of the Gamma distribution in the Gamma environment; as explained, these parameters determine the patient's probability of death as a function of time, modeled through the cumulative density function of the relevant distribution.

Finally, as previously explained, a and b determine the effect of the intervention (if any) by the doctor on the patient's probability of death. Where $p_t^{k,l}$ is the probability of death of a given patient at time t, as determined by k and l, performing the right intervention at t lowers the probability of death to $p_t^{k,l}/a$, and performing a wrong intervention at t raises that probability to $\left(b + p_t^{k,l} - 1\right)/b$.

A doctor intervenes at t if and only if at t her degree of belief in one of the hypotheses concerning α comes to exceed the threshold value of 0.9. If she intervenes, she makes the *right* intervention if what she believes in to a degree above the threshold is *true*, and a *wrong* intervention if it is *false*.

If the right intervention was made at time t, the doctor received the score of $1 - p_t^{k,l}/a$; if a wrong intervention was made, the doctor received the score of $1 - \left(b + p_t^{k,l} - 1\right)/b = \left(1 - p_t^{k,l}\right)/b$; if no hypothesis was assigned a probability above the threshold during the 100 time steps, and so no intervention was

made, the doctor received the score of $1 - p^{k,l}_{100}$. More exactly, given a patient characterized by $\langle \alpha, k, l, a, b \rangle$ and a doctor characterized by $\langle r, b, \Pr^{r,b} \rangle$, the doctor's score, s, for this patient equaled

$$s\big(\langle \alpha, k, l, a, b \rangle\big) \quad =$$

$$\begin{cases} 1 - p^{k,l}_t / a & \text{if } \exists h\, \exists t\, \forall t' < t\colon \ \big(\Pr^{r,b}_t(\alpha = h) > .9 \ \wedge \ \Pr^{r,b}_{t'}(\alpha = h) \leqslant .9 \ \wedge \ \alpha = h\big), \\ \big(1 - p^{k,l}_t\big) / b & \text{if } \exists h\, \exists t\, \forall t' < t\colon \ \big(\Pr^{r,b}_t(\alpha = h) > .9 \ \wedge \ \Pr^{r,b}_{t'}(\alpha = h) \leqslant .9 \ \wedge \ \alpha \neq h\big), \\ 1 - p^{k,l}_{100} & \text{otherwise.} \end{cases}$$

To repeat, after the doctor had treated her one hundred patients, her average score was calculated, and after this had been done for all 200 doctors, the one hundred doctors with highest scores were selected and copied and thereby came to constitute the population of the next generation, which went through the preceding selection process again, and so on, 250 times.

Computational details are not described here and are instead given in the Jupyter notebook that is part of the Supplementary Materials belonging to this chapter. The program for the simulations was written in Julia. In appendix E, I explain how the code for the simulations can be downloaded and used. Because Julia code reads almost like pseudocode, readers should be able to benefit from going through that code.

6.3.3 Results

I begin by describing the outcomes for the Weibull environment. Figure 6.4 gives a first impression of the kind of results obtained in this environment. It shows, for four randomly chosen simulations, how the population evolved through the 250 generations. One can think of each of the four plots as a series of 250 stacked histograms, where these histograms show how many tokens of each agent type were present in the corresponding generation.

We can immediately make a number of observations. Most notably, explanatory reasoning prevails in three of the four examples, in some form, and in the one example in which Bayesians (in the sense of doctors using Bayes's rule) are still in the running in the last generation, it appears that they are nevertheless losing the battle and might well have disappeared completely if we had not terminated the evolutionary process after 250 generations. And while there is no single explanation-based update rule that always trumps the others, EXPL users and Popperians do clearly better than users of Good's rule, who disappear quickly in all examples. Bayesians also disappear relatively

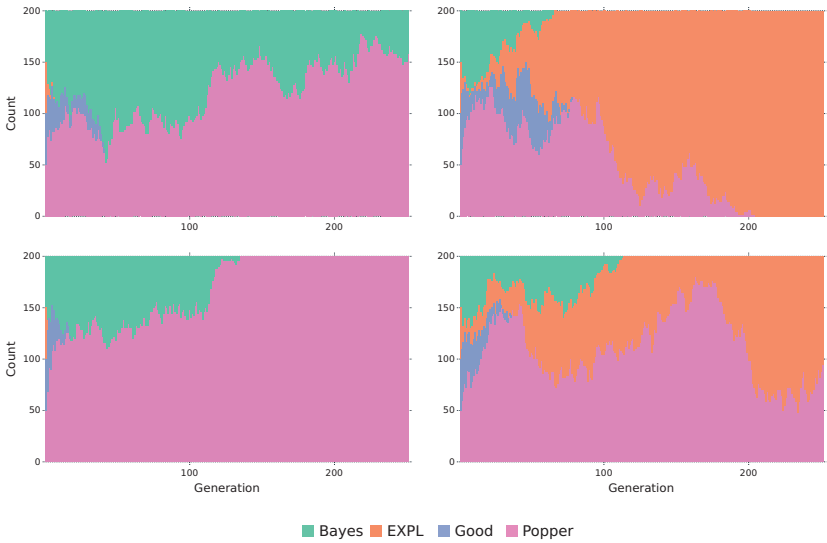

Figure 6.4: For the Weibull environment, counts of agent types per generation for four randomly chosen simulations.

quickly in three of the four examples. Also note that EXPL users appear to be critically endangered during the initial stages of all shown simulations. In the examples in which they perish, they do so *very* quickly, but where they make it through the initial stage, they become quite competitive, in one simulation even pushing out of competition all other types of agents.

As for the overall results, we first note that of all $50 \times 200 = 10{,}000$ agents in the last generations, 5,952 were Popperians, 2,246 were EXPL users, 1,554 were Bayesians, and 248 were users of Good's rule. Thus, users of Good's rule are indeed largely outcompeted into extinction. Furthermore, although among the explanation-based rules there is none that always comes to dominate the field, there is a clear sense in which Popper's rule is the winner, with EXPL being a distant second. The left panel of figure 6.5 plots the percentage of agents of each type of agent for every generation, averaged over all simulations. We note a strong tendency for the EXPL users to drop dramatically in number shortly after the start. We also see that their fate remains in serious jeopardy for some time after the start—if they do not perish completely, as the examples showed may happen. But when they manage to survive that

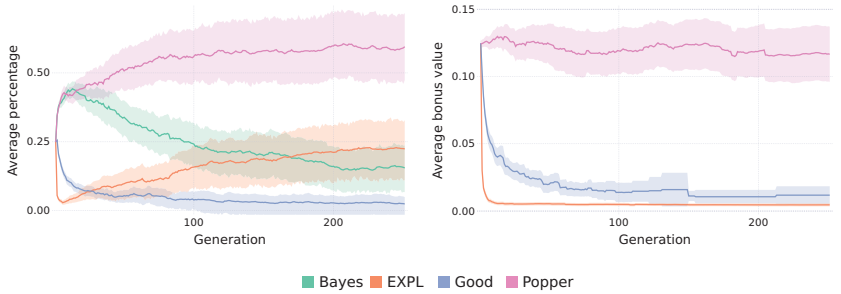

Figure 6.5: Percentage of agents per type of agents and per generation, averaged over the simulations for the Weibull environment (*left*), and mean bonus values per type of agents and per generation, averaged over the same simulations (*right*), both shown with 95 percent confidence bands.

critical phase, they are at least somewhat competitive against the Popperians. Bayesians appear competitive only for a very short while, after which they tend to fall back.

The right panel of figure 6.5 may suggest an explanation of why Popperians do better than EXPL users. This panel plots the mean bonus values in each generation for the three groups assigning such values. Because for the agents in all three groups the initial bonus values c were randomly drawn, with $c \sim \mathcal{U}(0, 0.25)$, all simulations start with an average bonus value of around 0.125 for each group. We see that the evolutionary process almost immediately drives down the average bonus value associated with the EXPL group to a level at which, as the plot also shows, it then stays for all further generations. Popperians, on the other hand, appear to get by with an average value close to 0.125: last-generation Popperians assign bonuses that are on average still close to the initial average value. So, Popperians may have an early advantage over EXPL users for which the latter never can really make up.

Turning to the Gamma environment, figure 6.6 shows how the evolutionary process unfolded in four randomly chosen simulations conducted in this environment. In these four examples convergence occurs very fast, in that one type of agents becomes dominant after fifty or even fewer generations. We further see that in these examples explanatory reasoning *always* prevails, with EXPL outcompeting all others in three of the four simulations. On the other hand, in one simulation EXPL users disappear again shortly after the

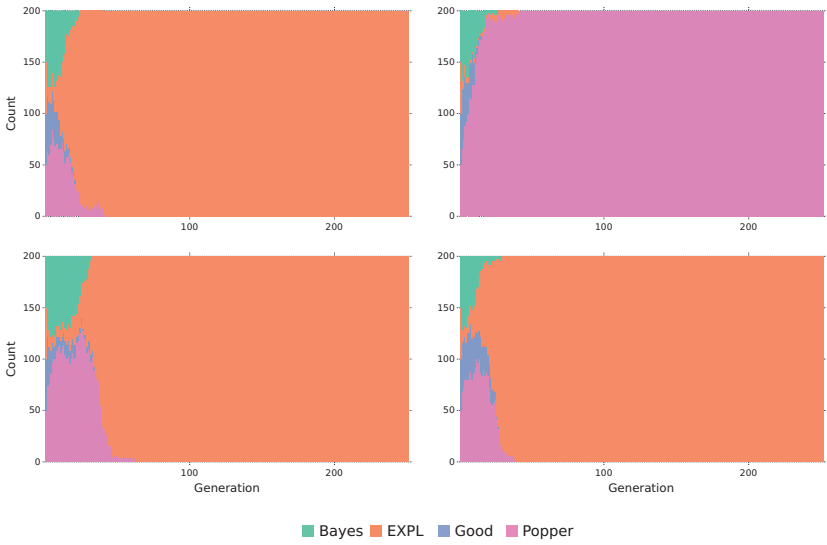

Figure 6.6: For the Gamma environment, counts of agent types per generation for four randomly chosen simulations.

start, leaving the field open to the Popperians. Even in the examples in which they end up as winners, EXPL users struggle at the beginning, similarly as in the previous simulations.

An analysis of the overall results shows that of the 10,000 agents in the last generations literally none are Bayesians or users of Good's rule. EXPL users form a clear majority, with a total presence of 7,554 agents in the last generations, the remaining 2,446 agents being Popperians. Thus, here EXPL is the clear winner, with Popper's rule being a distant second. The left panel of figure 6.7 gives a more complete picture, showing that the observations just made about the examples in figure 6.6 generalize: EXPL users appear to hang on by their fingertips for a short while, but if they are able to hold on, they tend to beat the entire competition. By contrast, although Bayesians and users of Good's rule do a bit better on average than EXPL users during the initial phase of the evolutionary process, they all go down the drain not long after.

In the right panel of figure 6.7 we see that the explanation of why EXPL users are endangered at first may again be that they have to drive their average

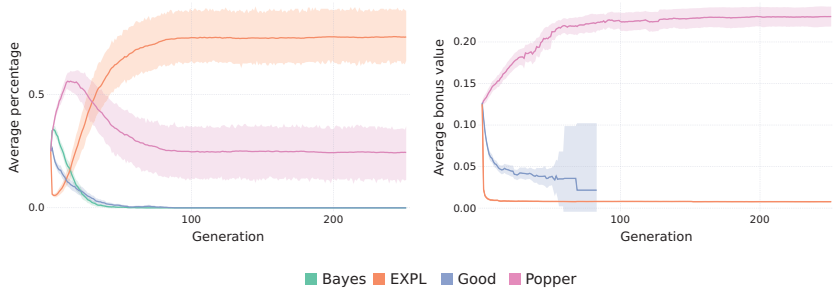

Figure 6.7: Percentage of agents per type of agents and per generation, averaged over the simulations for the Gamma environment (*left*), and mean bonus values per type of agents and per generation, averaged over the same simulations (*right*), both shown with 95 percent confidence bands.

bonus value down in order to become competitive. The same plot suggests an explanation of why in this environment Popperians do *not* outperform EXPL users: whereas in the Weibull environment the former's initial bonus values were fine from the start, at least on average, in the current environment their struggle for existence leads them to drive their bonus values up by, again on average, almost as much as the EXPL users drive theirs down. That the Popperians go through this process of adjusting their bonus values more slowly than the EXPL users may be why the former do worse, from an evolutionary viewpoint.

Next, let us compare the fitness levels—the scores on the fitness function, which measures the likelihood of patient survival—of the optimal solutions in the two environments with the corresponding fitness level of a Bayesian doctor. In the Weibull environment, the winner in the previous simulations was Popper's rule. Assuming that same environment, we run 10,000 simulations in each of which a Popperian treats one hundred patients, where the doctor always assumes the mean bonus value for last-generation Popperians in the previous simulations (which was .117, ± 0.021). We repeat this procedure with a Bayesian in place of a Popperian. And then we repeat all the foregoing but now assuming a Gamma environment and replacing the Popperian with an EXPL user, given that our simulations showed that EXPL users do best in this environment, and letting the EXPL user assign as a bonus to best explanations the mean bonus value for last-generation EXPL

users in the said simulations (.008, ± 0.001). In the new simulations in the Weibull environment, the Popperian doctor turns out to save an average of 84.1 (± 2.3) percent of the patients she treats, and a Bayesian doctor saves an average of 84.0 (± 2.3) percent of the patients she treats. Conducting a one-tailed t-test shows the difference to be significant ($t = 3.13, p < .001$), although the effect size is small (Cohen's $d = 0.04$). In the new simulations in the Gamma environment, we find that the EXPL user saves an average of 88.9 (±2.3) percent of her patients, and her Bayesian colleague saves an average of 88.1 (± 2.2) percent of patients. A one-tailed t-test shows the difference again to be significant ($t = 24.15, p < .0001$), although here, too, the effect size is small (Cohen's $d = 0.34$).

Although the small effect sizes could make these results look rather unexciting, consider that, on average in a Weibull environment, of every 10,000 patients admitted to the ICU, a doctor of the type that we found to do best in this environment using a bonus value that is optimal for the environment would save ten extra lives as compared to a Bayesian doctor. In a Gamma environment, the corresponding number of extra lives saved by the type of doctor best suited to operate in this environment is eighty. Numerically, these may not be large differences. Yet it is hard to imagine a doctor who would *not* in the environments at issue prefer Popper's rule and, respectively, EXPL (with suitable bonus values) over Bayes's rule—or to imagine a patient who would *not* prefer to be treated by a doctor using an optimal rule for the given environment.

Also note the following: we may conceive of the evolutionary algorithm that we used as only a means for answering the question of which rule would be ecologically most rational in the specific environments that we considered—the answer being that a doctor would be best off on average if she updated by Popper's rule in a Weibull environment and by EXPL in a Gamma environment, using the appropriate bonus value for the given rule in the given environment, and that therefore it is rational to use Popper's rule in a Weibull environment and EXPL in a Gamma environment. However, we can also think of the evolutionary algorithm as modeling somewhat realistically how, under evolutionary pressure, agents have come to prefer the use of this or that update rule in specific types of environment because of the relative competitive advantage that the rule offers. To reiterate a frequently made point, from an evolutionary perspective small differences may be as advantageous as larger ones: often enough, a prey animal can escape a predator by being just a *little*

faster than some other individuals in its group. That, I submit, best explains why Popperians do so much better than Bayesians in a Weibull setting, even though a Popperian doctor is only *slightly* "fitter" than her Bayesian colleague.

The broader evolutionary perspective further suggests that perhaps the most important advantage of the explanation-based update rules is that they offer an opportunity for adaptive learning—by increasing or decreasing the bonus for explanatory goodness—that Bayes's rule does not provide. How much a benefit this can be is especially clear if we focus on Bayesians versus EXPL users as represented in figure 6.5. Whereas the latter do quite a bit worse at the beginning, they can adapt to the needs of the context (as a group, not at the individual level, at which bonus values are fixed). As the figure shows, they were usually able to do this fast enough to avoid extinction and even to outnumber their Bayesian competitors. This is important, inasmuch as long-term survival is typically not just a matter of being well adapted to the environment in which one operates but also of being able to adapt quickly to a new environment.

The point of the foregoing is not that one should always prefer as an update rule some instance of Popper's rule in a Weibull environment and some instance of EXPL in a Gamma environment. Instead, it is to buttress Elqayam's previously quoted remark that it would be a mistake to expect a one-size-fits-all norm of rationality. According to the advocates of Bayes's rule, we are to follow it in each and every context. Not doing so would make us irrational, as the dynamic Dutch book and inaccuracy-minimization arguments are supposed to show. To refute this claim, it is enough to specify some context in which we are better off by following a non-Bayesian update rule. That leaves open the possibility that there are contexts in which Bayes's rule helps us achieve whatever it is we want to achieve more quickly or reliably or efficiently than would any other rule—which is fine, given that my aim was to show that in *some* contexts using a form of explanatory reasoning is ecologically more rational than using Bayes's rule and hence that Bayes's rule is not defensible as a *universally* valid principle of rationality; my aim was not to show that Bayes's rule is defensible under *no* circumstances, and that we should *always* adhere to some explanation-based update rule. Note that it is therefore not a flaw of the previous simulations that they relied on assumptions (about the characteristics of the patients, the time available for intervention, the threshold of certainty for action, and so on) that were to

some extent arbitrary.[9] That would be a problem only if my aim had been to argue for a replacement of Bayes's rule by a different supposedly universally valid update rule.

6.4 Concluding Remarks

Chapter 3 canvassed evidence indicating that the way people change their degrees of belief is influenced by explanatory factors and that this may cause them to violate Bayesian norms of reasoning. Why would they do that, if Bayes's rule has all the virtues its proponents allege it to have? The standard arguments in support of Bayesianism suggest an answer in terms of a bias, of something regrettable, even if perhaps not completely avoidable.

But upon close scrutiny in chapters 4 and 5, those arguments turned out to be fallacious. The key observation was that even granting that failure to minimize next-step inaccuracy and / or Dutch bookability are bad, nothing follows about abduction unless it has been ruled out that some compensation can be had in return. In this chapter, I have argued that people's tendency to rely on explanatory considerations in their belief updating can have clear advantages indeed and that, more generally, abductive reasoning is rational in some environments—rational in a sense that lets us focus on how well a cognitive strategy (e.g., an update rule) is adapted to the local environment in which it is deployed rather than on how well the strategy complies to certain internal standards such as consistency or probabilistic coherence.

The problem for the traditional Bayesian approach to updating, and to rationality in general, that we have aimed to highlight is not that it posits an ideal that we would not even know how to approximate. Instead the problem is that, at least in some situations, the use of Bayes's rule is simply *not* ideal. The computer simulations that were reported gave a straightforward example of this: in some contexts, use of Bayes's rule led to the survival of fewer patients, on average, than the use of a version of abductive reasoning. It was also seen that what helped some of those versions come out on top in the given contexts is that they are highly adaptable by having an adjustable parameter, which allows for contextual fine tuning. Bayes's rule lacks this functionality. More generally, because we inhabit an ever-changing world and can within a relatively short time span find ourselves in contexts that

9. However, see footnote 10 in chapter 4 on why .9 is a reasonable threshold value.

pose very different challenges, reasoning should not just be well adapted to the context that we are in at a given point in time; it should also be easily *adaptable* to whichever context we may, just moments later, find ourselves in (Schurz & Thorn, 2016). Insisting on the unique rationality of Bayes's rule would make us needlessly inflexible. If we are instead permitted to take on board instances of EXPL, Good's rule, Popper's rule, and other rules perhaps (including Bayes's rule), that would greatly facilitate quick adaptation, by switching from EXPL to Popper's rule, for instance, and / or by adjusting the explanation bonus.

It is fair to say that the approach to rationality that is prominent in much of the recent psychological literature, as we have been assuming in the preceding, has been concerned mainly with arguing for the ecological validity of so-called fast and frugal heuristics rather than that of the kind of higher-level cognitive principles that were compared in this chapter. A key insight of the work of Gigerenzer and his various collaborators is that the use of simple heuristics often leads to better outcomes than the use of more complicated principles of reasoning. However, nothing in Gigerenzer et al.'s work suggests that the notion of ecological rationality applies *only* to heuristics. Indeed, we showed in the previous section how a clear sense can be given to the notion of one update rule matching the environment better than another update rule.[10]

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

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

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

Another variation one could consider is to bring in the type of meta-inductive reasoners canvassed by Schurz, who keep track of the success rates of other reasoners and use that information to inform their own updating (see section 6.2). Schurz (2019, ch. 6) introduces a variety of meta-inductive strategies and discusses what difference they may make to predictive performance. It would be relatively straightforward to implement and compare those strategies in the context of our evolutionary computations.

Finally, although the simulations involved groups of agents, these agents went about their business all on their own and at no point were in touch with others. That is an important limitation. Alvin Goldman (1999) compellingly argued that much of traditional epistemology had been misguided in ignoring social interactions among people; that our flourishing qua epistemic agents, but also more broadly, can be understood only from a social perspective by recognizing the importance of our being embedded in a community of interacting agents collectively invested in the pursuit of truth. As Goldman pointed out, our knowledge would be a fraction of what it is had we had to explore the world all on our own. In a similar vein, Schurz (2019, p. 193) remarks that individual learning is far more costly than social learning and that "[m]any unsuccessful trial-and-error steps are involved in individual learning that can be avoided by just being informed about the results of these steps."[11] In this view, the traditional restriction to individual thinkers was bound to yield a distorted picture of human epistemology.[12] Therefore, the

11. On the importance of social learning, see also Boyd and Richerson (1985), Caporael (2001), Brewer (2004), Boyd, Richerson, and Henrich (2011), de Ridder (2014), and Oaksford and Hall (2016).

12. The importance of social learning in relation especially to the ecological conception of rationality has also been repeatedly emphasized by Gigerenzer (e.g., in his 2000).

next chapter again compares Bayes's rule and the various probabilistic versions of abduction that we have examined, now in a social setting in which agents' updates can be influenced not only by the evidence they receive directly from the world but also by the views of other agents in their community.