

CONCLUSIONS

This dissertation dealt with the construction of a rational model of inquiry. Within this model – based on Levi’s concept of *inductive expansion* – a research process can be described as a series of steps aimed at the acceptance of the best explanation available for the set of perplexities of a given agent at a given time. The rationality constraints of the model are weak, in the sense that all the model provides is a formal schema to be filled up with the agent’s epistemic values, beliefs and probabilities. The proposal can, in turn, be justified through its concrete applications. Hence, in the second part of the dissertation I sought to show that the account presented in the first part was well suited to help us understand the complexities of real cases in the history of science.

Part A comprised chapters 1, 2 and 3, and was aimed at providing a general account of IBE. Let me review the main ideas that appeared along these chapters.

Chapter 1 began by expounding what we might expect from an analysis of IBE, and sought to offer a justification for my choice of the topic. Then I turned to the historical roots of IBE, which, according to virtually all writers on the subject, should be traced back to the work of Charles S. Peirce. It is interesting to notice that most writers within the IBE literature identify the concept of IBE with Peirce’s notion of abduction. I tried to show that, if we are not careful, this identification might lead to confusions at many different levels. Besides a common tendency to conflate different epochs of Peirce’s work, one of the most unfortunate consequences of hastily linking IBE to Peirce’s texts is the lack of clarity as to whether an IBE involves, or does not

involve, acceptance of new ideas; related to this, another unfortunate consequence is the lack of clarity as to whether epistemic virtues (which Peirce explicitly mentions in his development of the notion of abduction) are meant to help us formulate hypotheses for the first time, or decide which hypotheses should be put to test, or perhaps help us choose a best explanation among prior explanatory (or potentially explanatory) statements.

In chapter 1 I also reviewed the current literature on IBE, ranging from works in the artificial intelligence field to Peter Lipton's well-known proposal (1991, 2004), as well as the many responses that appeared as a reaction to van Fraassen's (1989) claim that any IBE rule is bound to be incoherent. Finally, I also presented an outline of my own perspective, which was meant to be developed in chapters 2 and 3. In a nutshell, I suggested that we conceive of IBE as a process aimed at the acceptance of new hypotheses (though in the end acceptance might not occur). The many steps of this process can be grouped into two main stages. In the first stage an agent identifies a set of questions and potential answers to those questions. In the second stage the agent assesses the merits of each potential answer, and decides which one to incorporate to his or her set of full beliefs; suspensions of judgment (*i.e.*, disjunctions of hypotheses) also count as potential answers.

Chapter 2 addressed the first major moment of the IBE process. I began by presenting some tools drawn from formal belief revision theories, in order to achieve a suitable representation of an agent's epistemic state, as well as of the many ways in which an epistemic state can change. I submitted that the epistemic state of a given agent at a given time could be represented by a tuple $\langle L, S, K, C \rangle$ consisting of a

language L , a set of “meaningful” sentences S , a set of accepted sentences (or beliefs) K , and a set C of personal probability functions over the set of meaningful sentences. (Later, in chapter 3, I supplemented this tuple with additional elements). I argued in favor of the convenience of demanding that $K=Cn^*(K)$, where Cn^* is the restriction of the classical consequence operator to S ; S might not coincide with the inductive set of sentences of L .

The elements of this tuple can be modified in many different ways. For example, an agent’s belief set may incorporate new elements, or get rid of old ones. In addition, the set of meaningful statements can change, as well as the very representational language; I referred to these cases as “structural” shifts. Finally, the set of personal probability functions can also change; concerning this point I distinguished between Bayesian, anti-Bayesian, and non-Bayesian changes, depending on whether the shift of a given probability measure *followed* Bayesian conditionalization, *violated* Bayesian conditionalization, or whether Bayesian conditionalization did not apply, respectively.

The aforementioned representation of an agent’s epistemic state and of the many epistemic changes that such a state may undergo helped me describe the mechanism of the first stage of an IBE in a fairly precise manner. At the beginning of this stage, an agent’s epistemic state gives rise to a particular “interrogative state,” which includes a set of questions Q , and a set E of strongest consistent potential answers to all members of Q .

I contended that elements of Q did not need to be why-questions. This was not meant to deny the importance of why-questions in the context of IBE. As I argued in

chapter 3, why-questions might well have a central role to play (say, due to the fact that they might help agents achieve a more organized belief corpora) but the centrality of why-questions should not be seen as stemming from their ability to prompt a line of research in the first place; many other types of questions can do that as well.

As for set E , I required that it contain explanatory answers to all questions in Q . Members of E should be pairwise incompatible, and the agent should believe that exactly one of them is true. I did not define what an explanatory answer was. Rather, I argued that a satisfactory theory of explanation should focus on offering objective steps to determine when an agent is entitled to select a *best* explanation (or, eventually, a disjunction of best explanations) out of a previous batch of explanatory elements. By contrast, the theory should not tell an agent which hypotheses he or she should find explanatory in the first place; this amounts to a non-analyzable, ultimate fact. The resulting position can be described as a pragmatist conception of explanation, by analogy with the general layout of pragmatist epistemology. According to Peircean epistemology, agents should not seek justification for their prior beliefs, but only for the changes that occur in their sets of beliefs; similarly, according to a pragmatist conception of explanation, agents should not seek justification for finding certain statements explanatory, but for choosing to expand their belief sets with certain explanatory statements.

Elements of E may be put to a test; those which survive the testing constitute set U , an “ultimate partition” (for an agent and a time) relative to K and Q . No preferred answers are selected at this stage, so no members of U (nor disjunctions of members of U) are added to K yet. Still, many epistemic changes can take place in the

course of the process that starts with the identification of set Q and ends with the construction of U . In particular, the identification of suitable answers for given questions may prompt structural changes, which involve shifts in the set of the agent's meaningful statements; typically, structural changes also demand a number of non-Bayesian probability changes, as well as an expansion of K with new logical consequences.

Chapter 3 dealt with the selection context – the second major moment of an IBE process. Agents may choose to accept some of the hypotheses in U , or they may choose to suspend judgment among some, or all of them (there is also the logical possibility that they choose to accept a conjunction of rival hypotheses and thus contradict themselves, but it is clear that this option should be ruled out by any well-behaved model). This amounts to saying that an agent's set of *potential answers* is not actually U , but the set of all Boolean combinations of U , which I called U^A .

How should agents assess the options in U^A , and how should they decide which way to go – that is to say, which hypothesis, or disjunction of hypotheses, to accept, if any? (Incidentally, refusing to expand at all is equivalent to expanding with the disjunction of all elements of U). I suggested that we rely on Levi's brand of cognitive decision theory to answer these questions. Cognitive decision theory calculates the epistemic expected utility of different expansion strategies. In order to do this, it should be possible for agents to attribute both utilities and probabilities to each member of U^A ; hence, we need an adequate concept of epistemic utility. Levi had contended that epistemic utility functions should reflect a tradeoff between two main desiderata: avoiding error, and believing new things. According to Levi, the latter

could be expressed by means of a content function, based on a measure M that has the formal properties of a probability, such that the content of any element H of $U^A = 1 - M(H)$.

In chapter 3 I also sought to offer a way to fix M -values. To do this, I relied on the concept of *epistemic virtues* – more precisely, on features such as simplicity, unification power, fertility, testability, economy, and accuracy. Once again, agents are assumed to be able to perform a tradeoff among the many virtues that they deem relevant; the most virtuous hypothesis, for a given agent, has the lowest M -value, and thus the highest content. It should be noted that my view on epistemic virtues differs from that of most writers on the topic. I argued that, contrary to what seems to be a common assumption, the virtues just mentioned did not bear any clear relationship with truth or probability. They are not *truth-tracking*, so to speak. I contended that, at least with regard to the virtues mentioned above, the truth-tracking thesis required unsupported assumptions about the world, or involved a *petitio principii*.

The concept of unification power led me to consider a possible refinement of my proposal as to how epistemic states should be represented. I suggested that the tuple $\langle L, S, K, C \rangle$ be supplemented with a set of explanatory arguments W , and a set of patterns \mathbf{W} ; the more unifying a hypothesis is, the more it helps to achieve both a larger set W and a smaller set \mathbf{W} , with more stringent patterns. As a consequence of this analysis, I observed that we could rescue many of Kitcher's intuitions (as stated in his 1989) concerning what a theory of explanation should be about.

Suppose that an agent applies Levi's cognitive decision theory supplemented with my proposal on epistemic virtues to calculate the epistemic expected utility of

members of U^A . I have argued that, when the best element of U^A also happens to be a member of U , then that element should be interpreted as *the best explanation* for the set of perplexities that prompted the research, as far as the agent is concerned. In other words, I have argued that the overall explanatory force of members of U should be identified with their epistemic expected utility. On the other hand, if the best element of U^A is a disjunction of elements of U , I suggested that we call it *the best potential answer*, by the agent's lights. (Notice that a best explanation is then a limit case of a best potential answer – it is a best potential answer that has just one disjunct). The model, then, recommends that the agent adds the best element (say, H) to his or her set of beliefs K , such that the new belief set K' equals $Cn^*(K \cup \{H\})$. This change is typically accompanied by other shifts, such as a Bayesian probability update, and a change in sets W and \mathbf{W} .

The identification between the epistemic expected utility of elements of U and their explanatory force is actually a consequence of my view on explanation. I conceive of relative explanatory force as the ability a hypothesis has (comparatively speaking) to help agents enhance their overall understanding of the world – or, in other words, as the ability a hypothesis has to help agents build a more (or less) satisfactory world picture, by the agents' own lights. In this way, agents come to believe best explanations *because agents take them to be worth the risk*. Indeed, agents risk being wrong, but taking the risk may be rational if the gain in explanatory power is high enough.

This view departs from a common way of thinking about best explanations, according to which explanatory features of hypotheses point to the truth. According to

the more traditional picture, we are entitled to believe best explanations because explanatory force is the hallmark of truth. By way of contrast, within the model I favor, we are entitled to believe best explanations because explanatory force makes the risk of being wrong worth taking. (In the case of best potential answers that are not at the same time best explanations, we are entitled to believe them because, again, the benefits of accepting them outweigh the risks – even if those benefits are not best described as providing us with a more comfortable world picture.)

Part B comprised chapters 4, 5 and 6, and was devoted to concrete applications of the model. Clearly, the main goal of part B was to reconstruct actual IBE processes performed by actual historical agents, using the theoretical machinery of part A. But we could identify other important goals as well. For instance, part B was also meant to show how, by so reconstructing actual IBE processes with the aid of the model of part A, we might gain a deeper understanding of complex historical facts. In addition, the type of examples I chose were meant to show how it might be rational for different agents to arrive at opposite conclusions with regard to the same cluster of problems. In other words: I take it that sometimes it is intuitively rational for different agents to perform different IBEs on the basis of the same evidence; I sought to show how the model of part A enabled us to work out the details as to why this is so. Finally, the concrete applications were also meant to give us additional tools to appreciate that the model of part A had definite advantages over other epistemic proposals.

In chapter 4 I presented very briefly the main features of two historical cases (to be examined more carefully in chapters 5 and 6, respectively): the rise of

Mendelism, at the beginning of the twentieth century, and Theodore Avery's research on the nature of DNA, in the 1940s. The two examples could be described as turning points in the history of genetics (Avery's research being usually associated with the rise of molecular biology); they both involved the discussion of novel ideas (in the language of the first part, they both involved structural changes of some sort); and they both strike us as very complex cases from an epistemological point of view. On the other hand, in chapter 4 I argued that they did not resemble each other in the sense suggested by Stent (1972): unlike Mendel's, Avery's line of research was not particularly alien to his peers; lack of agreement does not automatically mean lack of understanding.

Chapter 5 dealt with the debate between William Bateson and Francis Weldon concerning Mendel's laws. I began by presenting Mendel's original paper, and then turned to its reception in the twentieth century. It is worth noticing that the chapter focused on *Mendelism* rather than on the real Mendel – particularly, it focused on *Bateson's* Mendel and his critics. I tried to identify both Weldon's and Bateson's belief sets, question sets, and ultimate partitions by the time they read Mendel's article, and then sought to reconstruct the evolution of their epistemic states in the years that followed; in Bateson's case, I extended my analysis to the first decades of the twentieth century. I submitted that we should think of Bateson's and Weldon's epistemic changes as the results of a series of IBE processes.

From the point of view of the historian of science, one of the most interesting features of the Bateson-Weldon debate is that the discussion often proceeded in ways that revealed the lack of sufficient common ground. The model of part A was very well

suited to register the discrepancies in shared assumptions and interests; paying attention to such differences helped me explain why the two characters acted rationally, in spite of achieving opposite conclusions.

Chapter 6 began by describing Avery, MacLeod, and McCarty's (1944) paper on bacterial transformation. Then I sought to identify the presuppositions of their research, the set of questions and answers that the members of the team suggested in the mid-forties, and the many possible IBE processes in which they took part. I also tried to clarify a number of misunderstandings about Avery's attitude towards his own research, as well as about the early reaction of the scientific community to the 1944 article. In particular, I argued that Avery's attempt to answer a question about the chemical composition of the transforming principle should not be conflated with an attempt to discover the *function* of such principle. Moreover, I pointed to the fact that Avery chose to ignore the second question because no clear way of settling it seemed available to him, and not because of a genuine lack of interest in finding an answer to it. A conflation between these two issues led many critics to unduly suggest a lack of understanding on the part of Avery of the real implications of his work.

Finally, I sought to identify the belief set, question set, and ultimate partition of Avery's most intransigent rival – Alfred Mirsky – by the time Hershey and Chase's (1952) paper on phage DNA was published. I also tried to provide a rationale for the peculiar result of Mirsky's IBE at the time – which could be characterized by a reluctance to accept that genes were DNA, on the face of the available evidence.

I have claimed that one of the goals of part B was to provide additional tools that might help us assess the IBE model of part A and see how it stands compared to other proposals. At least at first blush, the balance was positive: it turned out that the model was able to explain hypothesis acceptance even when such hypotheses did not bear substantially higher probability than their rivals; we have seen that the model made room for structural propositional shifts, as well as for non-Bayesian probability changes; finally, we have also seen that the model can be used to explain why different agents who infer competing hypotheses sometimes can be said to exhibit rational epistemic behavior.

The last feature may be taken to develop a number of basic Kuhnian intuitions – in particular, the intuition that sometimes disagreement about theories can be credited to the fact that different agents place different weights on different epistemic virtues (from Kuhn 1977). Unlike Kuhn's basic standpoint, however, the present model seeks to be *normative*: among other things, it makes concrete recommendations as to how agents should proceed. In any event, notice that the recommended paths will be different for each agent, depending on the concrete details of the agent's epistemic state – just like decision theory recommends different courses of action depending on the agent's utilities and probabilities. Thus, there are as many ways to fill up the details as researchers. Forcing the terms a bit, we could say that the present model attempts to recover some Kuhnian intuitions in a Levi-style framework.

As a final comment, let me consider very briefly two possible sources of concern. The first one is related to the fact that, at the time of assessing the epistemic behavior of a given agent, we are forced to provide *some* reconstruction of his or her

epistemic state, as well as of his or her belief dynamics. Then someone might protest that we can always accommodate things so that we never face an example of real irrationality – in which case the normative aspect of the model would simply vanish. Against this, I would like to point out that nothing in the model prevents us from using the given framework to actually distinguish rational from irrational epistemic behavior. In any event, I also believe that no abstract model should ever force us to violate our basic intuitions as to whether a concrete case in the history of science is an instance of good, or not-so-good, epistemic behavior. Rationality is in the eye of the beholder, so to speak. All a model should do is be flexible enough so as to embody and refine our prior intuitions in the best possible terms – as well as tell us how to proceed when no clear intuitions are at stake.

The second concern is related to the legitimacy (or lack thereof) of a normative model whose actual recommendations may differ from agent to agent, depending on the way the blanks are filled in. Someone might complain that this amounts to too much subjectivity for a theory of (best) explanation. In short, the model might be accused of fostering some form of relativism. In order to answer this concern notice that the task of providing rational reconstructions in no way commits us to share the view of those whose epistemic states we try to model. All we are required to do is assess whether such agents are objectively entitled to make certain inferential steps, taking their own assumptions into account. We can do so and still think that such assumptions are completely mistaken. In short, the model does not commit us to relativism. Quite the opposite, this model assumes that we, as interpreters, are always bound to speak from a certain point of view: namely, from our own.