paradigms allow for plausible criticism of the respective opponent on their own grounds. Seen from the critic's perspective, it is quite plausible to argue that a modification of the scientific paradigm in times of crisis is counter-productive, as it carries the risk of overlooking a solution that would satisfy the criteria set up by the old paradigm. Seen from the perspective of the new paradigm, sticking to the old criteria too long inhibits scientific progress by sticking to a misguided chimera of the static nature of scientific principles. All one can do in this case is assess the internal coherence and attractiveness of the old and new positions on their own terms and compare the two internal assessments on a more general – and therefore necessarily more vague – argumentative basis.

I argued in Section 1.1 that the traditional paradigm of theory assessment has run into a substantial crisis in the context of present-day fundamental physics. The next chapters will have a closer look at the question whether and to what extent the newly emerging paradigm of theory assessment in string physics and some other fields can provide a viable basis for overcoming that crisis. Let us begin by looking at the conceptual reasons string theorists rely on for believing in their theory. Later, an attempt will be made to put those reasons into a philosophical perspective.

1.3 Three contextual arguments for the viability of string theory

Why do string physicists invest trust in the viability of their theory? The main problem in this respect is addressed clearly in Penrose's cited text. Penrose raises one fundamental worry with regard to overly confident assessments of theories that have not been empirically confirmed: those assessments are always threatened by the possibility that other scientific explanations of the available data have been overlooked so far. Kyle Stanford has called this general problem of scientific reasoning the problem of unconceived alternatives (Stanford, 2001, 2006). Any reliable assessment of a theory's status on theoretical grounds must answer Penrose's worry by addressing the question of unconceived alternatives in some way. It shall be argued in the following that the current assessments of the status of string physics rely on a number of arguments which indeed amount to addressing that question.

Roughly, string theorists rely on two basic kinds of arguments when developing trust in their theory. Arguments which address structural characteristics of string theory itself shall be discussed in Part III of this book. Part I will focus on the other group of arguments, the contextual arguments which are based on general characteristics of the research process that leads towards string theory.

These arguments do not rely on any specific properties of the theory itself and are therefore of wider relevance. Part II will be devoted to demonstrating that they constitute an important part of theory assessment in fundamental physics in general.

Three main contextual reasons for the trust string theorists have in their theory may be distinguished. While all three arguments are "common lore" among string physicists, it is difficult to pinpoint a "locus classicus" for each of them. One can find a combination of all arguments in Chapter 1 of Polchinski (1998) and in Polchinski (1999). The first and third argument appear in Greene (1999, Chapter 1).

The plain no alternatives argument (NAA): string theorists tend to believe that their theory is the only viable option for constructing a unified theory of elementary particle interactions and gravity. It is important to understand the scope and the limits of that claim. String theory is not the only theory dealing with questions of quantum gravity. Various forms of canonical quantum gravity try to reconcile gravity with the elementary principles of quantum mechanics. They discuss the question of unification at an entirely different level than string theory, however. The latter stands in the tradition of the standard model of particle physics and is based on pivotal concepts such as non-abelian gauge theory, spontaneous symmetry breaking and renormalizability.[23] The goal of string theory is to reconcile gravity with these advanced and successful concepts of contemporary particle physics and therefore to provide a truly unified description of all natural forces. In this endeavor, the traditional investigations of canonical quantum gravity do not constitute alternatives, which leaves string theory as the only available way to go.[24] That is not to deny the relevance of the investigations of canonical quantum gravity. String theorists would just argue that, once the viable results of canonical quantum gravity are put into the context of contemporary particle physics, they will blend into the string theory research program.

Why is it so difficult to find a unified description of gravity and nuclear interactions? We have encountered the crucial problem for a unification of point-particle physics and gravity already in Section 1.1: quantum gravity is non-renormalizable within the traditional field-theoretical framework.

[23] A little more about those concepts will be said in Part II.

[24] There also exists a tradition of thought that questions the necessity of quantizing gravitation in a theory that gives a coherent description of quantum physics and gravitation. (Recent works are Wüthrich (2004), and Mattingly (2005).) Though some ideas concerning quantum theories of gravity without quantized gravity have been put forward, as yet none of them has been formulated in any detail, however. Like canonical quantum gravitation, those considerations at the present point address the reconciliation of gravity with basic quantum physics but do not offer concepts for a coherent integration of gravity and advanced particle physics.

Non-renormalizable quantum gravity cannot be considered viable at the Planck scale, however, the scale where the gravitational coupling becomes strong. Early attempts to solve this problem applied the traditional methods of gauge field theory and tried to deploy symmetries to cancel the infinities which arise in loop calculations and therefore make the theory finite. For some time, the concept of supergravity, which utilizes supersymmetry, looked like a promising candidate for carrying out this task, but eventually the appeal to symmetry principles was judged insufficient.[25] As it turns out, the remaining theoretical options are quite limited. One might venture into giving up some of the most fundamental pillars of present-day physics like causality or unitarity. Ideas in these directions have been considered, but did not lead to any convincing theoretical schemes. If one wants to retain these most fundamental principles, then, according to a wide consensus, there remains only one way to go: drop the idea of point particles, which univocally leads to string theory (see e.g. Polchinski, 1998, 1999).

A body of explicit analysis supports the notion that there may be no alternatives to string theory. Those considerations can be exemplified by an argument in Polchinski (1999). Polchinski starts with an innocent looking posit of a position–position uncertainty relation instead of the posit of extended elementary objects. He shows that the efforts to make that idea work eventually imply the very string theory he had set out to circumvent. Based on a number of arguments of a similar kind in conjunction with the fact that no alternatives have come up despite intense search, it may be suggested that string theory is the only option for finding a unification of all interactions within the framework of the long standing fundamental principles of physics.[26]

Naturally, the claim of alternatives does not remain uncontested. On what basis can we be confident that the scientists' lack of alternative ideas is not just a consequence of their limited creativity? Who can rule out that one of the most

[25] The case has not been closed entirely until this day. Recent analysis has suggested that there might be a finite ($N = 8$) supergravity after all (Bern, Dixon and Roiban, 2006). It still seems doubtful for a number of reasons, however, whether supergravity can be a fully consistent theory on its own.

[26] The history of science contains earlier claims of univocal inference from observation to the theoretical scheme. Newton's "deduction from the phenomena" has been taken up in Norton (1993, 1994), while Worrall (2000) has emphasized that deduction's dependence on prior assumptions. Newton's claim is based on the assertion of an immediate and intuitively comprehensible connection between observation and theoretical explanation. Compared to the case of Newton, the situation in string theory has decidedly shifted towards the assertion of a radical limitedness of options for mathematically consistent theory construction while the intuitive aspect of Newton's argument has been dropped. Whether this shift, in itself, enhances the authority of the string theoretical claim of no choice may be a matter of dispute. Part III of this book will demonstrate, however, that the string theoretical claim can be embedded in an entirely new and more powerful argumentative framework.

fundamental principles of physics indeed has to be jettisoned at this stage to describe nature correctly and that string theory is nothing more than a delusive "easy" way out that just does not accord with nature? Indeed, candidates for alternatives do sometimes appear where no alternatives had been in sight before and thus pose a constant threat to simple arguments of no alternatives.[27] It seems necessary to look for additional arguments for the viability of string theory which, rather than focussing on the scientists' difficulties to think of alternatives, are related to qualities of the theory itself.

Probably the most important argument of that kind is **the argument of unexpected explanatory coherence (UEA)**. It is widely held that a truly convincing confirmation of a scientific theory must be based on those of the theory's achievements which had not been foreseen at the time of its construction. Normally, this refers to empirical predictions which are later confirmed by experiment. However, there is an alternative. Sometimes, the introduction of a new theoretical principle surprisingly provides a more coherent theoretical picture after the principle's theoretical implications have been more fully understood. This kind of theoretical corroboration plays an important role in the case of string theory. Once the basic postulate of string physics has been stated, one observes a long sequence of unexpected deeper explanations of seemingly unconnected facts or theoretical concepts. Let us have a brief look at the most important examples.

String theory posits nothing more than the extendedness of elementary particles. The initial motivation for suggesting it as a fundamental theory of all interactions was to cure the renormalizability problems of quantum field theories that include gravity. Remarkably, string theory does not just provide a promising framework for quantum gravity but actually implies the existence of gravity. The gravitational field necessarily emerges as an oscillation mode of the string. String theory also implies that its low energy effective theory must be a Yang–Mills gauge theory, and it provides the basis for possible explanations of the unification of gauge couplings at the GUT-scale. The posit that was introduced as a means of joining two distinct and fairly complex theories, which had themselves been introduced due to specific empirical evidence, thus turns out not just to join them but to imply them.

[27] The case of new arguments for the finiteness of supergravity was mentioned in footnote 25. Another recent example for an interesting new perspective is Horava (2009), which proposes a scenario that solves the renormalizability problem of quantum gravity in a different way. It presents a theory that is non-relativistic at very short distances but approximates relativistic physics at longer distances. If consistent, such a scenario would constitute an alternative to string theory as a possible solution to the renormalizability problem of quantum gravity. The debate on the scenario's consistency, its promises and its limitations is ongoing at this point.

String theory also puts into a coherent perspective the concept of supersymmetry. Initially, interest in this concept was motivated primarily by the abstract mathematical question whether any generalization of the classical continuous symmetry groups was possible. As it turns out, supersymmetry is the maximal consistent solution in this respect. Soon after the construction of the first supersymmetric toy-model, it became clear that the implementation of supersymmetry as a local gauge symmetry (i.e. supergravity) had the potential to provide a coherent quantum field theoretical perspective on gravity and its interaction particle, the graviton. In the context of string theory, on the other hand, it had been realized early on that a string theory that involves fermions must necessarily be locally supersymmetric.[28] The question of the maximal continuous symmetry group, the quest to integrate the graviton naturally into the field theoretical particle structure, and the attempts to formulate a consistent theory of extended elementary objects thus conspicuously blend into one coherent whole.

A problem that arises when general relativity goes quantum is black hole entropy. The necessity to attribute an entropy proportional to the area of its event horizon to the black hole in order to preserve the global viability of the laws of thermodynamics was already understood in the 1970s (Bekenstein, 1973). The area law of black hole entropy was merely an ad hoc posit, however, lacking any deeper structural understanding. In the 1990s it turned out that some special cases of supersymmetric black holes allow for a string theoretical description where the black hole entropy can be understood (producing the right numerical factors) in terms of the number of degrees of freedom of the string theoretical system (Strominger and Vafa, 1996). Thus, string physics provides a structural understanding of black hole entropy.

All of these explanations represent the extendedness of particles as a feature that seems intricately linked with the phenomenon of gravity and much more adequate than the idea of point particles for a coherent overall understanding of the interface between gravity and microscopic interactions. The subtle coherence of the implications of the extendedness of elementary objects could not have been foreseen at the time when the principle was first suggested. It would look like a miracle if all these instances of delicate coherence arose in the context of a principle that was entirely misguided.

To what extent is it justified to have serious confidence in the viability of string theory's phenomenological predictions on the basis of the presented

[28] World sheet supersymmetry of a string that includes fermions was discovered by Gervais and Sakita (1971). A string theory that shows local target space supersymmetry was finally formulated by Green and Schwarz (1984).

"non-empirical no miracles argument"? There do exist cases in the history of science where the inference from a concept's success to its viability was invalidated by the fact that just one aspect of the concept was responsible for the success while important parts of the concept were misguided. A prominent example would be the ether theories whose success was based on the viability of the wave equation but whose core concept, the ether, had to be dropped eventually. In the case of string theory it is difficult to imagine anything of that kind, since the concept is based on one simple and entirely structural posit which would seem impossible to reduce without taking it back altogether. It must be considered a genuine possibility, however, that a deeper, more fundamental principle than string theory itself could be responsible for unexpected explanatory interconnections without implying string theory itself. Still, as broad a spectrum of unexpected explanatory interconnections as is encountered in the case of string theory, in conjunction with the apparent difficulty to come up with consistent alternatives to the theory, may seem difficult to reconcile with the idea that those explanatory successes do not hinge on the theory itself. One thus may have the impression that the argument of unexpected explanatory interconnections has some strength but may not feel confident enough to rely on it without further analysis. Given that there is no option at this point for carrying out empirical tests of string theory itself, it would thus seem important at least to find a possibility of checking the general validity of the non-empirical arguments for string theory's viability on an empirical basis. The third argument is based exactly on this kind of empirical test at a meta-level.

The meta-inductive argument from the success of other theories in the research program (MIA): most string theorists, at any rate those of the first generation, are mainly educated in traditional particle physics. Their scientific perspective is based on the tremendous predictive success of the particle physics standard model. The latter was created for solving technical problems related to the structuring of the available empirical data (in particular, the problem of making nuclear interactions renormalizable) and it predicted a whole new world of new particle phenomena without initially having direct empirical confirmation. Just like in the case of string theory, it turned out that none of the alternatives to the standard model that physicists could think of was satisfactory at a theoretical level. In addition, surprising explanatory interconnections emerged. (For example, the distinction between a confining interaction like strong interaction and the non-confining electromagnetic interaction could be explained as a natural consequence of the difference between a non-abelian and an abelian interaction structure.) Given the entirely theoretical motives for its creation, the lack of satisfactory alternatives and the emergence of unexpected explanatory interconnections, the standard model can be called a direct precursor of string theory.

Indeed, string theorists view their own endeavor as a natural continuation of the successful particle physics research program. The fact that the standard model theory was at the end impressively confirmed by experiment conveys a specific message to particle physicists: if you knock on all doors you can think of and precisely one of them opens, the chances are good that you are on the right track. Scientists working on unifying gauge field theory and gravity have thought about all currently conceivable options, including those which drop fundamental physical principles. The fact that exactly one approach has gained momentum suggests that the principles of theory selection which have been successfully applied during the development of the standard model are still working.

It is important to emphasize that MIA relies on empirical tests of other theories and thereby in a significant sense resembles the process of theory confirmation by empirical data. The level of reasoning, however, differs from that chosen in the classical case of theory confirmation. In the present context, the empirically testable prediction is placed at the meta-level of the conceptualization of predictive success. We do not test the scientific theory that predicts the collected empirical data but rather a meta-level statement. The following hypothesis is formulated: scientific theories which are developed in the research program of high energy physics in order to solve a substantial conceptual problem, which seem to be without conceptual alternative and which show a significant level of unexpected internal coherence tend to be empirically successful once they can be tested by experiment. This statement is argued for based on past empirical data (in our case, largely data from the standard model of particle physics and from some earlier instances of microphysics) and can be empirically tested by future data whenever any predictions which were extracted from theories in high energy physics along the lines defined above are up to empirical testing.

All data that can be collected within the high energy physics research program thus constitute relevant empirical tests of the viability of MIA and thereby have implications for the status of other theories in the field which are considered likely viable based on that hypothesis. Any experiment that confirms predictions whose viabilities have been considered likely based on purely theoretical reasoning would improve the status of scientific theories of similar status in the research field. On the other hand, the status of non-empirical theory evaluation and therefore also the status of theories believed in on its basis would seriously suffer if predictions that are strongly supported theoretically were empirically refuted. Trust in string physics on that basis is influenced by new empirical data even when that data does not represent a test of string theory itself.

Present-day high energy physics provides two excellent examples for the described mechanism. At the LHC, two theories were and still are tested which, before the start of the experiment, were considered probably viable (to different degrees) based on theoretical reasoning. First, the canonical understanding of high energy physics implied that the LHC was likely to find a Higgs particle.[29] The Higgs mechanism constituted the only known method of producing the observed masses of elementary particles in a gauge theoretical framework. Since gauge field theory had proved highly successful in the standard model, where all empirical predictions other than the Higgs particle had already been empirically confirmed, and since the theoretical context of mass creation by spontaneous symmetry breaking[30] in gauge field theory was well understood, physicists would have been profoundly surprised if the Higgs particle had not been found at the LHC. (A closer analysis of that case will be carried out in Chapter 4.) The actual discovery of the Higgs particle in summer 2012 clearly constitutes an example of an eventual empirical confirmation of a theory that had been conjectured and taken to be probably viable on theoretical grounds. Thereby, it constitutes a confirmation not just of the Higgs theory but also of the meta-level hypothesis that the involved theoretical strategies of theory assessment are viable. In the given case, the general belief in the existence of the Higgs particle was so strong that its discovery did not alter overall perspectives too much. A failure to find the Higgs, however, would have constituted a serious blow not just to the current understanding of the standard model but to the status of non-empirical theory assessment in general. String physicists in that case would have had to answer the question on what basis they could be confident to get the basic idea of string theory right on entirely theoretical grounds if high energy physics could not even correctly predict the Higgs particle. Since the principle of the viability of non-empirical theory assessment can only be of a statistical nature, it could not be refuted by individual counter instances. Trust in string physics would have seriously suffered, however, if a prediction as well trusted as the one regarding the Higgs particle had failed.

The second theory that may get empirically confirmed at the LHC is low energy supersymmetry. The situation there is a little different than in the case of the Higgs particle. As will be discussed in more detail in Section 4.2, some conceptual arguments do hint towards low energy supersymmetry. The cogency of those arguments, however, is less clear than in the Higgs case or in the case of string theory. Many string physicists would argue that there remain more conceptual options for avoiding low energy supersymmetry than for avoiding

[29] Either as a fundamental or as a composite particle.
[30] See Section 4.1 for a brief explanation of spontaneous symmetry breaking.

string theory. Thus, if low energy supersymmetry was not found at the LHC that would obviously not help the trust in string theory but the damage would be limited. On the other hand, if both the Higgs particle and low energy supersymmetry were found, that would be taken as significant support for the reliability of theory generation based on conceptual reasoning. Trust in non-empirical theory assessment would get considerably strengthened, which in turn would enhance the trust in string theory.[31]

Having discussed the three most important arguments of non-empirical theory assessment, it is now possible to be more specific about what is meant by "non-empirical" in the given context. The term "non-empirical" clearly does not imply that no observation or no empirical data has entered the argument. As we have seen, "non-empirical" theory assessment does rely on observations about the research process, the performance of scientists looking for alternative theories or the success of theories in the research field. What distinguishes empirical from non-empirical evidence in the given sense is the following. Empirical evidence for a theory consists of data of a kind that can be predicted by the theory assessed on its basis. Non-empirical evidence, to the contrary is evidence of a different kind, which cannot be possibly predicted by the theory in question. No scientific theory can predict that scientists will not find other theories which solve the same scientific problem (the observation that enters NAA). Nor can a scientific theory possibly predict that other theories which are developed within the research field independently from the given theory tend to get empirically confirmed (the observation that enters MIA). Non-empirical evidence for a theory thus is evidence that supports a theory even though the theory does not predict the evidence.

In conjunction, the three presented arguments of non-empirical theory assessment lay the foundations for trust in string theory. All three reasons have precursors in earlier scientific theories but arguably appear in string theory in a particularly strong form. The complete lack of empirical evidence in the given case leads to a situation where non-empirical theory assessment is solely responsible for the status attributed to string theory and therefore is of special importance.

[31] The described support for string theory at the meta-level must be distinguished from reasoning at the theoretical "ground level." The discovery of supersymmetry would support string theory at the theoretical ground level as well, since string theory predicts supersymmetry (though it does not imply low energy supersymmetry).

2

The conceptual framework

Section 1.3 presented the reasons for trust in string theory from the perspective of physicists. A number of crucial questions remain open at that level, however. Is there a common basis for the three arguments for having trust in string theory? Can those arguments be called genuine scientific reasoning? Do they amount to theory confirmation? Finally, can we find a philosophical background story behind the described rise of non-empirical theory assessment?

In order to deal with those questions, we have to take a step back and define the philosophical context within which the presented situation evolves. I will begin by offering a more substantial philosophical definition of the canonical paradigm of theory assessment that motivates much of the criticism against string theorists' trust in their theory. Based on that characterization, I will then introduce the core philosophical concept which shall be argued to underlie the ongoing paradigm shift with respect to theory assessment. This concept, which will determine much of the later analysis in this book, goes under the name "assessment of scientific underdetermination."

2.1 The classical paradigm of theory assessment

A minimal and fairly uncontroversial first description of theory assessment in science may be given in the following way. Scientific theories make predictions that can be tested by collecting empirical data. If the collected data turns out to be in agreement with the theory's predictions, this enhances the scientists' trust in the theory's viability (in whatever way we may want to specify that trust in terms of the attribution of truth, empirical adequacy, the reliability of its empirical predictions or other concepts). If the data contradicts the predictions, that lowers trust in the theory correspondingly.

Most philosophers of science, in particular those who are guided by the example of physics, take one crucial step beyond the characterization given above. They insist that confirmation by empirical data is the only way a scientific theory can acquire the status of an acknowledged and well-established scientific theory. According to this understanding, scientists are only entitled to believe in their theory if that theory has been empirically confirmed by empirical data. The same basic idea may be expressed in terms of the conception of knowledge: a scientific theory can only constitute knowledge about the world if it has been confirmed by empirical data. Based on the classical definition of knowledge as justified true belief, the statement means that the scientist cannot get justification for her belief in a theory without empirical confirmation. In terms of an alternative definition of knowledge as belief that has been produced by a reliable cognitive process, empirical confirmation constitutes a necessary element of any reliable cognitive process that leads towards establishing a scientific theory.

The exclusive role of empirical testing in scientific theory assessment is evident in the most influential schemas of the scientific process. According to hypothetico-deductivism, scientists produce a new theory based on prior knowledge of empirical data by an act of creative speculation. The resulting theory at first has the status of an untested hypothesis. Empirical predictions are then deduced from that hypothesis. The hypothesis can be tested only by confronting its predictions with actual empirical data. Only if the theory's predictions get repeatedly confirmed empirically, the hypothesis can become a well-established and trusted theory. On the other hand, disagreements between predictions and empirical data weaken and eventually refute the hypothesis. The hypothetico-deductivist thus adamantly holds that confirmation and refutation by empirical data are the only possible ways to determine a theory's viability.

Bayesianism, which currently constitutes the most popular formalization of scientific theory confirmation, conveys a similar albeit not identical message. The (subjective) Bayesian assumes that the scientist attributes a probability of truth to each scientific theory. Bayes' theorem,

$$P(\mathrm{T}|\mathrm{E}) = \frac{P(\mathrm{E}|\mathrm{T})}{P(\mathrm{E})} P(\mathrm{T}),$$

determines how new empirical evidence E alters the probability of the theory's truth. $P(\mathrm{T})$ denotes the prior probability of the truth of hypothesis H before the empirical data E has been considered; $P(\mathrm{E})$ is the probability of the empirical data E disregarding H; $P(\mathrm{H}|\mathrm{E})$ is the probability of H when E has been taken into account; and $P(\mathrm{E}|\mathrm{T})$ is the probability of E given that H is true. Empirical data constitutes confirmation of theory H if

$$P(\mathrm{T}|\mathrm{E}) > P(\mathrm{T}).$$

Probabilities according to this schema can only be assigned based on prior probabilities attributed to theories, which themselves must either be posited or assessed based on another Bayesian procedure. No algorithm exists for determining initial probabilities. They are subjectively chosen by the scientist based on her general assessments, her experience, personal preferences, etc. According to the Bayesian, this irreducible subjective element can be reconciled with the notion of generally agreed upon scientific knowledge because repeated empirical tests tend to create a convergence behavior for probabilities that are derived from a wide range of early prior probabilities. Empirical data that agree with the theory's predictions quickly raise low prior probabilities $P(\mathrm{T})$ to levels close to those reached when starting from rather high $P(\mathrm{T})$s. Empirical data that contradict the theory's predictions, on the other hand, quickly reduce high initial probabilities $P(\mathrm{T})$. Probabilities in mature science are thus taken to be fairly well decoupled from the prior probabilities attributed at early stages of the scientific development.

Unlike hypothetico-deductivism, Bayesianism does not exclude the attribution of very high initial probabilities to scientific theories. Bayesianism is coherent with a situation where the scientist takes a new theory to be very probably true for some reason without having tested it empirically. However, the Bayesian approach strongly emphasizes that the scientific process of theory assessment starts in earnest with the theory's confrontation with empirical data E. Only then do the probabilities that initially have an entirely subjective character, and may be chosen quite differently by individual scientists, converge and thus start constituting reliable scientific judgements. The empirical testing of the theory by new data is what the Bayesian considers the genuinely interesting part of the process of theory assessment.

Bayesianism as well as hypothetico-deductivism implicitly make a second assumption regarding the scientific process. They assume that scientific theories which are to be tested empirically acquire a fairly complete state within a reasonable time span. Only once they have assumed that state is it possible to deduce predictions from the theory along the lines suggested by hypothetico-deductivism; and only sufficiently complete theories can be reasonably evaluated in a Bayesian way. Both conceptions of the scientific process thus distinguish a period of theory construction from a distinct period of empirical testing.

Joining the considerations presented above, we can define the contours of a prevalent overall understanding of the scientific process that is expressed in hypothetico-deductivism as well as in "canonical" Bayesianism. A notion of

science along these lines is shared by most philosophers of science today and is taken to be a reliable guide for scientific activity by scientists themselves in modern physics and related fields. The scientific research process is thereby understood in terms of a dichotomy between theoretical conceptualization and empirically based theory assessment. On the one hand, in the realm of theory building scientists create theories in order to structure and describe the available empirical data and, eventually, predict new data and new phenomena; on the other hand, the process of empirical observation and experimentation first inspires theory building and then, based on new data, provides confirmation or refutation of the developed theoretical statements. A theoretical conception has to meet certain structural preconditions in order to count as a candidate for a viable scientific theory: it has to constitute a largely complete and internally coherent theoretical structure; and it has to offer quantitative predictions the empirical side can aim to test. If a conception fulfils those conditions and thus is being awarded a "candidate status" for becoming a scientifically viable theory, empirical testing constitutes the only genuinely scientific method for determining that theory's status. As long as the theory's core predictions have not been empirically confirmed, the theory remains a scientific speculation. Only after empirical confirmation has been forthcoming may the theory be accepted as viable scientific knowledge.

I will call this understanding of the scientific process the classical scientific paradigm. What the classical scientific paradigm unequivocally denies is the possibility that a process of rational analysis on its own can offer an alternative strategy to empirical testing for turning a scientific hypothesis into a well-established and well-trusted theory. Purely theoretical considerations like those pertaining to a theory's simplicity, beauty or apparent cogency may contribute to the scientist's subjective trust in a theory's prospects of being viable. They are acknowledged as being of potential auxiliary value to the researcher who tries to decide which way to go before empirical tests have been carried out. They are not considered an objective and independent factor, however, in determining a given theory's scientific viability.

The assertion of a univocal primacy of empirical data in theory assessment has been questioned and toned down in a number of ways by various philosophers. Two prominent examples may be mentioned.

Thomas S. Kuhn (1962) emphasized that the interpretation of empirical data always happens within some paradigm based on scientific background convictions and thus can never provide an objective account of its own significance. It may happen that aspects of empirical data which are considered important evidence under one paradigm are considered largely irrelevant under another. Kuhn and his followers deny that the confirmation value of empirical data for a

theory can be determined on a basis that is independent of the paradigm represented by that theory. This implies that the choice of a scientific paradigm cannot be based on empirical data alone but must be based on a broader form of deliberation whose outcome is not univocally determined by rational analysis and which involves conceptual, theoretical and even social elements.[1]

While Kuhn thus alters the understanding of the role of empirical data in the research process, he does not question that the individual scientist who works within some paradigm (either the one she inherited from her peers or a new one she is testing) looks for confirmation of her theory based on empirical data according to the procedures outlined in previous paragraphs. Though paradigm choice according to Kuhn is not decided based on empirical confirmation or refutation alone, the testing of a theory within its paradigm does proceed based on empirical testing nevertheless. In the eyes of the working scientist, theories thus assume a status of being well established based on empirical testing. In this respect, Kuhn has no quarrel with hypothetico-deductivism.

Larry Laudan's analysis of theory assessment in his book *Progress and its Problems* (Laudan, 1977) moves a little further. Laudan aims at modifying the understanding of theory assessment and theory choice without putting a strong emphasis on the Kuhnian distinction between paradigm change and normal science. Thereby, non-empirical theory assessment turns into a canonical element of theory assessment at all times and is not confined to the revolutionary phases of paradigm change. Laudan argues that the dominating influence of the empiricist paradigm has blinded philosophers of science to the simple fact that scientists judge their theories' status and viability not merely based on empirical merits but also and substantially based on theoretical qualities. What makes this claim significant is Laudan's insistence on the equal status of both strategies of theory assessment. According to Laudan, a careful historical investigation of the process of theory selection in science does not justify asserting a univocal hierarchy between empirical and theoretical strategies of theory assessment. Theoretical strategies do not play a strictly auxiliary role that would restrict their influence to cases when the empirical verdict is not yet clear. Rather, theoretical arguments under certain conditions can be considered more important than empirical ones and may overrule empirically based preferences.

[1] Unlike some of his followers, Kuhn himself does not deny that the eventual prevalence of one theory and the demise of its contenders can be rationally explained. According to Kuhn, a period of normal science starts once one paradigm has turned out so successful that its superiority over its known contenders can no longer be rationally denied even from those contenders' perspectives. Only in periods of crisis and paradigm change does the identification of the most successful paradigm actually depend on the paradigm on which the assessment is based (see Hoyningen-Huene, 1993).

The main claim to be pursued in this book shares the basic sentiment behind Laudan's analysis. It shall be argued that the classical paradigm of theory assessment grossly underrates that assessment's theoretical aspect. However, the analysis will choose a significantly different point of departure than Kuhn or Laudan. The conception to be promoted is theory assessment based on assessments of limitations to scientific underdetermination, which does not play a significant role in Kuhn's or Laudan's work. In order to introduce the idea of scientific underdetermination, let us return once more to the foundations of the classical paradigm of theory assessment.

2.2 Scientific underdetermination

Why does it make sense for scientists to believe in the empirical predictions of theories which have already been empirically confirmed and refuse to believe in predictions made by empirically unconfirmed theories? In order to answer that question, we must first introduce the important distinction between prediction based on straightforward induction and prediction of genuinely new phenomena. The first kind of prediction may be exemplified by the expectation that the scattering process between two hard and massive macroscopic objects that has so often been found to proceed in agreement with the Newtonian principles of mechanics will again adhere to those principles the next time we observe it. A classic example of the second kind of prediction would be Poisson's inference from a wave theory of light to a bright spot at the center of a shadow cast on a screen by a round object when hit by light that has passed first through a narrow hole.[2]

Scientists are willing to believe scientific predictions of the first kind because they are willing to rely on the principle of induction. The extension of the applicability of simple enumerative induction to many precisely specified scientific contexts is taken to be one central reason for the success of the scientific principle.

If a scientist constructs a theory, however, that (a) fits the available data and (b) predicts new phenomena which have not yet been observed, her trust in the actual existence of the newly predicted phenomena is restrained by one crucial consideration: other, so far unknown scientific theories may exist which fit the present data equally well but predict different new phenomena. In other words,

[2] The borders between the two kinds of predictions cannot be univocally drawn. The two kinds of predictions rather constitute ideal types segmenting a continuous spectrum rather than univocally distinguishable groups. Still, the distinction is helpful for conveying the rationale behind the scientists' dealings with their theories.

Underdetermination by...	(a) all possible evidence:	(b) the available evidence:
(1) Logically:		Hume, problem of induction; Quine-Duhem thesis
(2) Ampliatively:	Quine ('Reasons for Indeterminacy of Translation'); van Fraassen (The Scientific Image')	Sklar, Stanford: transient underdetermination; SCIENTIFIC UNDERDETERMINATION

Figure 2.1 Underdetermination of scientific theory building.

scientific theory building is expected to be significantly underdetermined by the currently available empirical data. I will call this presumption the principle of "scientific underdetermination."

Scientific underdetermination has to be distinguished from two types of underdetermination which figure most prominently in philosophy.[3] In order to distinguish the different kinds of the underdetermination of theory building it may be helpful to draw a systematic picture (see Figure 2.1).

Philosophers of science speak either about (1) underdetermination considering all logical possibilities or about (2) underdetermination under some general assumptions which are taken to be constitutive of all viable scientific research. Such general assumptions, which are called "the ampliative rules of scientific method" in Laudan (1996), may include a commitment to an intuitive form of the principle of induction, some kind of Ockham's razor and the exclusion of non-integrated ad-hoc explanations of individual events. These assumptions may differ from one scientific field to the other. In each field they determine what a scientist in the field would consider a legitimate scientific statement. Orthogonal to this distinction, it is important to differentiate between (a) underdetermination by all possible empirical data and (b) underdetermination by the currently available data.

The concepts of underdetermination which have played the most substantial roles in the philosophy of science up to now were exemplifications of versions (1b) and (2a). Hume's claim that a given data set does not logically imply any future events refers to the available evidence and asks for logical possibilities. Thereby, it constitutes the classic example of a claim of type (1b) underdetermination. The

[3] Using the name "scientific underdetermination" is not meant to suggest that other types of underdetermination are of no or lesser interest to a philosophical analysis of the scientific process. The name has been chosen based on the point made above that scientific underdetermination is the kind of underdetermination most relevant to the acting scientist herself.

Quine–Duhem thesis, which holds that any individual statement can be made compatible with any new empirical data by making appropriate changes within the wider conceptual framework, would be another example for a claim of (1b) underdetermination. Quine's statements regarding the underdetermination of scientific theory building in Quine (1970, 1975) deal with the options for constructing scientifically viable and empirically equivalent theories. Therefore, they exemplify (2a). A wide range of recent philosophical thoughts about underdetermination also falls into this category. Laudan and Leplin (1991), van Fraassen (1980) or Sklar (2000) would be prominent examples.

Version (2b) arguably is the one most relevant to the scientist who searches for new theories. In order to be able to develop theoretical schemes at all, the scientist must take for granted the validity of a basic principle of induction, the existence of a coherent theoretical scheme capable of describing the phenomena in question and some vague assumptions about the universality, predictive power and lack of ad-hoc-ness of that scientific scheme. She thus must take for granted a framework of ampliative rules of the scientific method. Science thus is concerned with underdetermination of type (2). Moreover, the scientist aims at producing successful predictions based on the currently available data. The scientific search for new theories must be based on the empirical evidence currently available, which implies that the kind of underdetermination of interest to the scientist in this context is of type (b). The claim of (2b) underdetermination in a certain field at a given time asserts that it would be possible to build several or many distinct theories which qualify as scientific and fit the empirical data available in that field at the given time. Since these alternative theories are merely required to coincide with respect to the presently available data, they may well offer different predictions of future empirical data which can be tested by future experiments. It is type (2b) underdetermination due to the existence of such empirically distinguishable theories which will be of primary interest in the following analysis.

Type (2b) underdetermination so far has played a less prominent role in the philosophy of science than (1b) or (2a). It is by no means unknown in the philosophical literature, however. Lawrence Sklar (1975) may have been the first to discuss it as a distinct form of underdetermination under the name transient underdetermination. Kyle Stanford (2001, 2006) has recently emphasized the importance of transient underdetermination for the scientific realism debate. The reason why the term "transient underdetermination" is not adopted in this book has to do with the profound differences between the perspective on underdetermination chosen by Sklar and Stanford and the one to be developed in this book. Sklar and Stanford stay within the canonical paradigm of theory assessment and understand underdetermination as something that must be

established by actually finding alternative theories. Once alternative theories have been found, underdetermination with respect to those theories can be removed by carrying out empirical tests which decide between these alternative theories. Seen from this perspective, the use of the term transient underdetermination for type (2b) underdetermination appears natural. The core claim of this book, to the contrary, will be that the degree of type (2b) underdetermination can be assessed without knowing the alternative theories. Therefore it becomes important to understand the degree of underdetermination in terms of the number of *possible* alternatives, irrespective of the question whether those alternatives are known or not. "Transient underdetermination" from this perspective would suggest that all *possible* alternative theories of today can be removed by experiment in the future so that, at some stage, there will remain no alternative to the surviving scientific theory any more at all. But this is obviously not what Sklar, Stanford or I want to assert by claiming type (2b) underdetermination.[4] Therefore, using the term "transient underdetermination" for type (2b) underdetermination would be highly misleading in the context of this book, which is why I use the term "scientific underdetermination" instead.

The distinction introduced in Section 2.1 between scientific speculations and scientific knowledge can now be motivated by referring to the principle of scientific underdetermination. Well-established scientific theories are those whose distinctive predictions[5] have been experimentally well tested and confirmed in a certain regime. The general viability of the theory's predictions in that regime is considered a matter of inductive inference.[6] Speculative theories, on the other hand, are those whose distinctive predictions have not yet been experimentally confirmed. Even if a speculative theory fits the currently available experimental data, its distinctive predictions might well be false due to the scientific underdetermination principle.

Scientists and philosophers of science endorse the principle of scientific underdetermination for a number of reasons. First, scientific underdetermination is supported by the vaguely instrumentalist spirit inherent in the prevalent understanding of the scientific process. According to this understanding, the scientist builds theoretical structures which reflect the regularities observed in nature up to some precision and tunes the involved free parameters in order to fit the quantitative details of observation. The successful construction of a suitable theory for a

[4] In fact, the question whether or not all possible alternatives can be excluded so that only one theory remains will be the main topic of Part III of this book.

[5] That is, predictions, whose experimental confirmation would be direct empirical support for the novel theoretical claims of the theory.

[6] A scientist's formulation of the notion of well-established theories can be found e.g. in Weinberg (2001).

significant and repeatedly observable regularity that characterizes the world is assumed to be just a matter of the scientist's creativity and diligence. If it is always possible to find one suitable scientific theory however, it seems natural to assume that there can be others as well. Different choices of theoretical structure must be expected to exist which have coinciding empirical implications up to some precision in the observed regime if their respective free parameters are fixed accordingly. The principle of underdetermination follows from this.

Furthermore, the assumption of scientific underdetermination constitutes a pivotal element of the modern conception of scientific progress. If science proceeds, as emphasized e.g. by Kuhn (1962) or Laudan (1981), via a succession of conceptually different theories, all future theories in that sequence must be alternative theories which fit the present data and therefore exemplify scientific underdetermination.[7] Theoretical progress without scientific underdetermination, to the contrary, would have to be entirely cumulative.[8]

In addition, there are many concrete examples of scientific contexts where the available empirical data is known to allow a wide spectrum of conceptualizations with substantially different empirical predictions with respect to new data. In particle physics, to give one example, we know that many different extensions of the particle physics standard model are compatible with the presently available data.

The importance of the principle of scientific underdetermination thus is well established. Scientific underdetermination cannot be taken to be entirely unconstrained, however. If it were, that is if all imaginable regularity patterns of empirical data could be fitted by fully satisfactory scientific theories, no correct predictions of new phenomena could ever be expected to occur. One would rather be led towards understanding the specific predictions offered by the presently available theory as one accidental "pick" among an infinite number of theoretically viable options. In this light, it would seem entirely unreasonable to take those predictions seriously. It is a fact, however, that successful predictions of new phenomena do happen frequently in advanced science. Scientific underdetermination thus must be assumed to be limited in some way.

[7] It should be emphasized that the last statement would lack distinctive meaning if based on the most radical reading of Kuhn's incommensurability thesis. The statement relies on the assumption that adherents of the successive theories referred to can find a consensus with respect to the scientific characterization of the collected empirical data. In the context of particle physics, which shall be analyzed in the present work, this assumption clearly seems justified as adherents of all existing particle physical theories share the same understanding of the implications of specific particle experiments for theory building.

[8] The scientific underdetermination principle is closely related to the pessimistic meta-induction of Laudan (1981) but does not share the latter's anti-realist claims. It reflects an ontologically neutral assessment of the status of scientific theories that is fairly uncontroversial in recent science and philosophy of science.

Does science have to care about the nature and extent of those limitations? Is the assessment of those limitations part of the scientific process? Does it constitute a theoretical element of scientific theory assessment? As we have argued above, the classical empirical paradigm of theory assessment does not allow for any such role. It will be the task of the present book to argue otherwise.

3

The assessment of scientific underdetermination in string theory

3.1 Connections between theory assessment and scientific underdetermination

It may seem surprising at first glance that scientific underdetermination should play any role at all in assessing string theory. Section 2.3 introduced scientific underdetermination as underdetermination of scientific theory building by the available empirical data. String theory, however, has remained entirely unconfirmed by empirical data up to now. The question of underdetermination would arise more naturally if we were dealing with a theory that has found some empirical confirmation and would ask the question how many alternatives could be constructed in agreement with the available empirical data.

In order to understand the rationale of the argument in the given case, one must remember that string theory has been developed based on and informed by empirical data. It has been constructed in order to provide a universal theory that underlies, and has as its approximations in certain regimes, two theories (general relativity and gauge field theory) which have been empirically confirmed. At present, the data that confirms those theories does not constitute empirical evidence for string theory because it cannot be predicted based on the current understanding of string theory. However, since string theory conceptually relies on the given empirical data, the conditions for applying assessments of scientific underdetermination are fulfilled. A complex web of posits, conjectures and physical analysis lies between the available empirical data and the theoretical concept of string theory. As discussed in Section 1.2, it is the threat of scientific underdetermination that normally prevents trust in theoretical speculations whose connections to available empirical data are that distant. In this light, it is natural to expect that only an assessment of scientific underdetermination can increase the trust in such theories. Claims of limitations to scientific underdetermination have the potential of strengthening the inferential connection

between the empirical data that has motivated the theory's construction and the theory itself. In the following we will argue that the three arguments for the viability of string theory presented in Chapter 1 can be understood in this sense.

Let us first have another look at NAA, the no alternatives argument. The observation that no equally satisfying solutions to a given scientific problem have been discovered by scientists even after a long and careful search for alternatives can be interpreted in two different ways. First, one could conclude that the depth and scope of the scientific analysis at the time just have not been sufficient for finding the appropriate alternatives. Some alternatives within the chosen framework may have been overlooked or some of the foundational postulates which have defined the framework for the overall search may require modifications in order to allow alternative solutions. That conclusion would offer no reason for believing that the lack of known alternatives to the available theory has any implications for that theory's chances of being empirically viable. Limitations to human intellectual capacities provide no good argument for the viability of the theory scientists came up with. The situation may also be interpreted in a different way, however: one might conjecture a connection between the spectrum of theories the scientists came up with and the spectrum of all possible scientific theories that fit the available data. The observer who chooses this second path takes the fact that scientists have problems finding alternatives as a sign that not too many alternatives are possible in principle. In other words, she concludes that scientific underdetermination is significantly limited. Based on the additional assumption that the phenomena in question can be characterized by a coherent scientific theory at all, the conjecture of significant limitations to scientific underdetermination can then enhance the trust in the available theory's viability: if a viable scientific theory exists and only very few scientific theories can be built in agreement with the available data, the chances are good that the theory actually developed by scientists is viable. The step from an observation about the present human perspective to a conclusion regarding the overall spectrum of possible scientific thinking is by no means trivial and raises deep philosophical questions. Still, it is a necessary precondition for using the argument of no choice as an argument for a theory's viability.

UEA, the argument of unexpected explanatory coherence, is related to an assumption of limitations to underdetermination as well, albeit in a less direct way. The initial observation in the given case does not characterize the human activity of finding and developing alternative theories like in the case of the no alternatives argument. It rather deals with properties of the theory itself. The reasoning relies on the observation that some explanatory connections provided by the theory were not aimed at during theory construction but have emerged

after closer analysis of the theory's structure. The argument thereby mirrors the canonical reasoning for a theory's viability based on novel empirical confirmation. The importance of distinguishing between unexpected explanations and intended ones can be argued for in analogy with the distinction between novel data and data that has entered the process of theory construction. If physicists were searching for a microscopic explanation of black hole entropy and came up with a theory whose only merit was to offer such an explanation, it would not make sense to take that merit as a strong indication for the theory's validity. Since that was what people were searching for, no one could be surprised that they eventually came up with a theory of that kind. If, like in the case of string physics, the theory was developed for other reasons, however, the gratuitously received extra merit is of significance. It gives the impression that physicists are on the right track.

But why is it that surprising explanatory power can reasonably be taken to enhance a theory's chances of being viable? Naturally, a theory that solves a higher number of conceptual problems of predecessor theories can be considered more likely to be viable. This does not account for the additional aspect, though, that the theory was not constructed to provide those solutions. In order to understand the extra value of surprising explanatory power, let us first imagine a scientist who approaches the problem (p1) how to unify gravity and gauge field theory under the assumption that solutions to that problem are abundant. Let us further assume that she takes solutions to other seemingly independent problems like the problem (p2) how to explain the scale hierarchy between electroweak scale and Planck scale or the problem (p3) how to acquire a microscopic understanding of black hole entropy to be abundant as well. Finding a theory that solves one of those problems from her perspective does not imply that this theory is physically viable. The physically viable solutions (i.e. the ones whose empirical core predictions are correct) may well be among the many other possible theories. Now let us assume that this physicist finds string theory to be a solution to the unification problem (p1). Without additional information, the described physicist does not have any reason to expect that string theory will also be a solution for problems (p2) and (p3). If the physicist adheres to a general form of scientific optimism and assumes that there is an empirically fully adequate scientific theory that covers the research field to which all three problems belong, she must expect that among the large number of solutions to each of the problems there will be at least one which is empirically fully adequate and therefore solves all three problems at once. There may even be a number of other theories which solve all three problems as well. It is to be expected, however, that a vast number of theories, probably most theories which offer a solution to any of the three

problems at all, will only solve one of the problems. The given physicist has no reason to believe in advance that string theory is not an element of the latter group.

Now compare this with the perspective of another string theorist who is a strict supporter of NAA and believes that there are no possible alternatives to string theory. Based on the principle of scientific optimism introduced above, string theory thus must be taken to be the true theory. Since the given string physicist must expect that the true theory solves all problems, she is forced to expect that string theory will solve all problems. In other words, the expectation of explanatory interconnections the theory was not constructed to provide is natural for someone who believes in the most rigid limitations to scientific underdetermination while it is unnatural for someone who believes in the unconstrained abundance of theoretical solutions. Finding unexpected explanatory power therefore supports the conjecture of limitations to underdetermination. UEA can strengthen the case for NAA.

Even in conjunction, however, the two arguments NAA and UEA are not conclusive. As argued in Chapter 1, unexpected explanatory power could have other reasons than a lack of alternatives to the corresponding theory. It could arise due to so far insufficiently understood theoretical interconnections at a more fundamental level, of which the theory in question is just one exemplification among many others. If so, the observed unexpected explanatory power would indicate the viability of that underlying more fundamental principle rather than the viability of the theory itself. In order to distinguish between the two cases, it would be helpful to get a better grasp of the actual chances of empirical success of theories which show a strong pattern of unexpected explanatory success.

MIA, the meta-inductive argument from the success of other theories in the research program, provides information to that end in an intricate way. As discussed above, MIA is an empirical argument. The empirical data deployed does not serve as an empirical test of the theory under consideration, however. It is used at a meta-level, providing an empirical test of the strategies on non-empirical theory assessment. If those strategies are found to be regularly successful, they become more trustworthy. The empirical observations that provide the basis for MIA thereby increase the trust in so far empirically unconfirmed scientific theories which are supported by the given strategies.

It is of crucial importance that MIA conveys a message of limitations to underdetermination already on its own terms. The line of reasoning to that end is similar to the one presented above in the context of UEA and has already been briefly addressed at the end of Chapter 2. Let us, in analogy to the discussion of argument UEA, consider an observer who assumes that there are many different

possible theories which can account for the empirical data available at some point. Now, let us look at predictions of novel phenomena offered by those theories. (In the context of particle physics, such predictions will primarily be concerned with new elementary particles and their properties.) Without further assumptions, the given observer must expect that the various theories will offer many different predictions with regard to the next generation of empirical tests. On these grounds, chances seem rather small that the theory developed by scientists in order to account for the available data will offer correct predictions with respect to the upcoming empirical tests. The assumption of limitations to underdetermination, on the other hand, can provide an explanation of predictive success: the fewer the scientist's options for constructing scientific theories that fit the available data and offer different predictions regarding the upcoming experimental tests, the better are her chances of selecting one of those theories which give correct predictions of the results of those tests.

The question remains whether predictive success could be explained in other ways as well which do not rely on claims of limitations to scientific underdetermination. Some potential alternative explanations come to mind. For example, one might assume that specific properties like simplicity or beauty guide scientists towards finding predictively successful theories. If scientists were indeed striving for simple and beautiful theories and if there were a deep connection between simplicity or beauty and predictive success, that connection might serve as an explanation for the predictive success of the theories scientists develop.

Would this constitute a viable alternative explanation of predictive success? One may have the general objection that the suggested explanation suffers from the notorious problems to define simplicity and beauty in the given context. It is by no means self-evident that the theory development in fundamental physics or any other field moves towards "simpler" theories in any sense that goes beyond higher universality. If no development towards simplicity in a significant and independent sense can be confirmed, however, it would seem untenable that higher simplicity is correlated with higher predictive accuracy. In the context of fundamental physics, it is not clear either, whether it makes sense to attribute to the scientists the conceptual freedom to choose between simpler and less simple or beautiful and less beautiful theories. More often than not, scientists must be content with finding any coherent solution to their conceptual problems at all.

Even if we assumed that definitions of simplicity and beauty can be provided in a meaningful way and do play a significant role in theory building, however, closer inspection reveals that the corresponding explanation of predictive success does not constitute an *alternative* to limitations to scientific underdetermination but must rely on the assumption of limitations to scientific underdetermination itself.

Let us assume that scientists have found a theory with a given degree of simplicity and beauty. Explaining the predictive success of that theory on that basis must rely on the assumption that only few alternative theories with similar or higher degrees of simplicity or beauty exist. A certain degree of beauty or simplicity thus is turned into a condition which has to be fulfilled by those theories with regard to which limitations to underdetermination are assessed. In other words, invoking simplicity, beauty or similar criteria which are expected to be fulfilled by predictively successful theories merely changes the framework of conditions within which assessments of scientific underdetermination are carried out. An explanation of predictive success that works without any reference to limitations to scientific underdetermination would have to abstain from imposing any such criteria and assume a human capacity to find predictively successful theories by means which transcend the analysis of the available empirical data and the structural properties of potential theories. It would thus amount to positing a kind of magical "truth detector" that finds the true theory just because it is true. This, however, seems alien to scientific reasoning. Whenever scientists do manage to develop various conceptually satisfactory solutions to a given scientific problem, all those theories must be considered as viable candidates for the valid solution of that problem. No magical "truth detectors" can be deployed in order to find the true solution.

Assuming limitations to scientific underdetermination therefore indeed looks like the only satisfactory explanation of the fact that scientists regularly find theories which are predictively successful. Inference to the best explanation then can lead from the observation of regular predictive success in a scientific field towards the conjecture that scientific underdetermination is limited in that field. Any instance where empirical predictions get confirmed by empirical tests thus can be understood as an indication of limitations to underdetermination in the given regime.

The next step of MIA consists in making the meta-inductive inference that regular predictive success in a research field justifies the assumption that future predictions of a similar kind will be correct as well. To be applicable, the inference must rely on a reasonable understanding as to what can count as a prediction of a similar kind. Obviously, regular predictive success in some research field does not imply that every theory that is constructed in the field in agreement with the known data is likely to be predictively successful. Any inference from one theory's predictive success to the probability of another theory's empirical viability must be based on reasons to believe that the new theory is in a significant way comparable to the earlier ones. This is where limitations to underdetermination become a crucial bond between the three arguments of non-empirical theory assessment. We have noted that predictive

success can be linked to limitations to scientific underdetermination. We have further noted that NAA and UEA offer independent reasons for believing in limitations to underdetermination in specific contexts. Above, we have seen that MIA can be understood in terms of limitations to scientific underdetermination as well. Therefore, we have good reasons for taking the conditions corresponding to NAA and UEA to be good criteria for selecting the group of theories within which we can make legitimate inferences based on MIA. If novel predictive success occurs regularly when theories fulfil criteria for NAA and UEA, it is justified to expect with some confidence that it will occur with respect to another theory that satisfies conditions for NAA and UEA as well.

The interrelated web of reasoning presented above consists of arguments that mutually support and strengthen each other and provide a machinery of theory assessment that is based on empirical data. NAA and UEA provide criteria for the inference carried out in MIA. MIA, on the other hand, may be understood in terms of an empirical test of the criteria defined by arguments NAA and UEA. If theories which were (or could have been) considered probably viable based on NAA and UEA turn out to be empirically successful in many instances in a scientific field, that can be taken as empirical corroboration of the viability of the criteria provided by NAA and UEA.

The question now arises: can this kind of reasoning be acknowledged as a critical scientific method which has some legitimacy for providing a foundation for theory assessment? Any scientific strategy of theory evaluation must specify the circumstances under which its criteria of scientific success speak against a theory's viability. Moreover, there must be specifiable possible circumstances under which the status of the strategy itself as a viable strategy of theory assessment weakens on its own grounds. In the given case, it would clearly be insufficient to specify empirical data that refutes the theory under scrutiny. Since the entire enterprise has been started in order to allow for trustworthy theory assessment in the absence of empirical data predicted by the theory in question, the scenarios under which the assessment speaks against the theory's viability must be of a non-empirical nature as well.

Indeed, it is possible to list a number of scenarios which would significantly reduce or even nullify the trust in string theory based on purely theoretical evaluation criteria. One can also find scenarios in which the strategies of non-empirical theory assessment would lose much of their appeal. Let us look at such scenarios one by one.

NAA relies on the observation that no conceptually equally satisfactory alternatives to string theory are found. It would have to be withdrawn whenever a conceptually satisfactory alternative to string theory was in fact discovered. Any such alternative would thus significantly reduce the trust in string theory.

UEA would lose strength if some of the interconnections in question turned out to be based on deeper and more general patterns of reasoning which were not univocally related to string theory. Let us assume that string theory provides an explanation for some structural characteristics we find in physics which did not have a satisfactory explanation before. It may happen that, after a while, we understand that this explanation is based on a deeper physical principle X that could be extracted without any reference to string physics. String theory in that case just would have been the context where we first understood a principle X that was in fact far more general than the specific theory. In that case, we must not be surprised that string theory provided the given explanation since, due to X, any alternative theory would have provided that explanation as well. Therefore, we could not use the fact that string theory provides the explanation in question as an argument for the viability of string theory but only for the viability of more general assumptions which generate principle X. In fact, considerations along these lines are being discussed in string physics and contribute to a better understanding of the significance of the argument of unexpected explanatory interconnections. For example, it has been argued that microscopic calculations of black hole entropy can be calculated based on general principles without using a string theoretical framework (see Strominger, 1998, and Carlip, 2008).

MIA can lose power based on new empirical data regarding other theories. As argued already in Section 1.3, trust in string theory would be substantially reduced if theories which seemed to have no alternatives turned out to make false predictions. If no Higgs particle had been found during the LHC experiments at CERN, many observers would have raised the question how confident one could be regarding the remote claims of string theory if one could not even correctly predict the existence of the Higgs particle.

On a more general basis, theoretical arguments in favor of string theory could also be weakened by other developments which would cast doubt on the chances of success of the research program. For example, an improved theoretical understanding of string physics might reveal theoretical weaknesses which change the theoretical assessment of the theory's chances of being fully consistent and thus physically viable. Furthermore, an interruption of theoretical progress over a long period of time could raise doubts whether a more complete theoretical understanding of string physics were attainable at all.

New information can also reduce the trust in the method of non-empirical theory assessment itself. Generally, each instance where a theory has been taken to be probably viable based on non-empirical theory assessment and then lost its trustworthiness again due to new evidence, be that evidence empirical or non-empirical, weakens the status of non-empirical theory assessment in the very same way that it is strengthened by instances which show its consistency or predictive power.

To conclude, we face a situation where theories can be supported as well as weakened based on non-empirical evidence. The same kinds of analysis which can provide theory confirmation without finding data that reproduces the theory's predictions, can, under different circumstances, amount to the disconfirmation of the theory without having found data that contradicts the theory's predictions.[1] Therefore, it seems indeed possible to grant non-empirical theory assessment the status of viable scientific reasoning.

The skeptic regarding the significance of non-empirical theory assessment might still have one fundamental objection. Nothing of what has been said so far, she might say, has dissolved the fundamental difference between the empirical confirmation of a scientific theory and non-empirical theory assessment. Even if one may find reasons for giving some significance to the latter, the fundamental epistemic difference between observing a phenomenon on the one hand and inferring a statement's validity based on circumstantial evidence on the other must be upheld. As a matter of principle, trust in an empirically unconfirmed theory can never be compared to trust in a theory that has been empirically confirmed.

The answer to this objection is twofold. First, the presented arguments do not suggest that non-empirical theory assessment can replace empirical confirmation or assume the very same status within the structure of scientific reasoning. Since non-empirical theory assessment crucially relies on empirical confirmation within the research field, it is in an important sense secondary to the latter.

Second, however, the difference in epistemic status between empirical confirmation and non-empirical theory assessment is less fundamental than one might assume at first glance. The substantially stronger position of empirically confirmed theories compared to theories supported by non-empirical reasoning is not an immediate consequence of the scientific method but rather a non-trivial empirical fact about the world we face.

We often trust empirically well-confirmed theories to the extent that we are willing to bet our lives on their viability: we use aeroplanes or bridges based on our understanding that they have been built according to well-confirmed scientific theories. In fact, we would call irrational anyone who refused to do so. A person that refuses to step on a bridge because of doubts regarding the involved physical laws would be considered profoundly strange.

To the contrary, few people would be ready to trust a theory to the same extent based on non-empirical theory assessment. But let us imagine, for a moment,

[1] In fact, the latter scenario is quite uncontroversial in some cases; finding out that a theory constitutes just one out of a million theoretical options to fit the available data makes its viability quite nearly as unlikely as finding data that contradicts its predictions.

some "strange world" that is very different from the world we live in. Physicists in this "strange world" have developed theories for 10 000 years and every single theory they have developed in a mathematically coherent way without being able to find an alternative eventually has turned out to be empirically viable. In that world, NAA (the argument that the observation that scientists have not found an alternative to a theory indicates a theory's viability) would be considered every bit as trustworthy as standard inductive inference based on empirically well-confirmed scientific theories. It would be obvious in this "strange world" that physicists have the capacity to exhaustively consider possible alternatives even if no one had come up with an explanation how they do it. A person that doubted a theory that was confirmed by NAA would seem quite as weird in the "strange world" as a person that does not believe in the stability of bridges does in ours.

Obviously, we do not live in a "strange world." In our world, physicists do have problems finding existing alternative theories and the confirmation value of non-empirical evidence is more precarious and difficult to assess – which is why the debate about the significance of non-empirical theory assessment is highly non-trivial and by no means stupid. But the mechanisms which can lead us towards acknowledging the confirmation value of non-empirical evidence remain the same in principle in our world as in the "strange world." If science enters a phase where such arguments work better than in earlier periods or other contexts of scientific reasoning, it is therefore rational and necessary to modify the status of non-empirical theory assessment accordingly. Non-empirical theory assessment in our world obviously is more cumbersome, less stable and less precise than straightforward theory assessment based on empirical testing. Assessing the degree to which it is, however, constitutes a crucial scientific task.

3.2 The framework for claims of limitations to scientific underdetermination

Any statement asserting that only few scientific theories can be developed which fit a given data set must be based on a definition of what counts as a scientific theory. Without specifying a sufficiently rigorous framework within which we count possible theories, we cannot meaningfully claim that the number of those theories is limited. If one would not take for granted a commitment to the validity of inductive reasoning, to take the most obvious example, theory building would be free to predict any pattern of future events imaginable and thus could not be taken to be limited in any meaningful sense. In Chapter 2, we established a framework for statements on underdetermination by restricting

our analysis to ampliative scientific reasoning. The term "ampliative" denoted the assumption of a certain set of conditions which have to be fulfilled by a theory in order to be called scientific. Statements on limitations to scientific underdetermination must be understood within this framework of presumptions. Two important questions arise here. What is the basis for restricting the considerations on limitations to underdetermination to the framework of ampliative scientific reasoning? And to what extent is it necessary to specify the conditions of ampliative reasoning in order to provide a solid foundation for statements on scientific underdetermination?

In order to get a clearer understanding of those questions, let us for the moment forget about the ampliative conditions of scientific reasoning mentioned above and start with the most basic statement. A claim of limitations to underdetermination must be made within some framework that defines which kinds of theories are addressed by the claim. The selection of an adequate framework then must satisfy two conditions which drag in opposite directions. On the one hand, it must be sufficiently strong for allowing significant statements on limitations to underdetermination. On the other hand, it must be sufficiently wide for allowing high confidence that a viable scientific description of the phenomenology to be predicted can be found within its limits.

If the second condition were not met, we would have a nice statement on limitations to underdetermination but could not use it for the purpose of establishing that a theory has good chances of being viable. In order to understand the problem, let us imagine that we choose a specific set of general physical principles – for example the principles of gauge field theory in particle physics – as a framework for a statement on limitations to underdetermination. Since we understand the gauge principle fairly well, this would be a convenient choice for making strong statements on limitations to underdetermination. However, the question would arise as to how probable it is that the next generation of empirical data can in fact be described successfully by a gauge field theory. This new question would have to be assessed by starting a new level of assessment of underdetermination: we would have to assess the number of scientific theories which are not gauge theories, reproduce the available data and give predictions regarding new data which differ from those of the gauge theories. If there are many of them, we would have to consider it likely that the viable theory will not be a gauge theory. After all, we have no good reason to believe that the scientific principles we have happened to develop at this point are more likely viable than other principles which offer equally satisfactory descriptions of the available data. In order to carry out that second level assessment, however, we would have to introduce a second level of scientific principles that provides the framework for that assessment. Those more general

scientific principles may be questioned again and we face a threat of an infinite regress. Using known scientific concepts as frameworks for statements on limitations to underdetermination thus does not help in establishing any assessment of a theory's viability.

The only way to avoid this problem consists in retreating to a framework of presumptions that is general enough for justifying trust in its capacity of allowing viable descriptions without further assessments of underdetermination at a meta-level. Using ampliative criteria of scientific reasoning looks like a natural choice since it conflates trust in the framework's legitimacy with trust in the success of scientific reasoning. The scientist thus can say that, qua being a scientist, she believes in the viability of the scientific method and takes it for granted. It is a presumption closely related to the belief in the general viability of an inductive principle: if a strong inductive inference that relies on our present body of scientific knowledge fails, the scientist takes it for granted that this can be explained based on a new scientific theory.

The problem may look less threatening now, but it has not vanished. The question remains as to how stable the present criteria of scientific reasoning must be. In fact, we know of a number of criteria which would have been taken to be essential elements of any legitimate scientific theory by most scientists at an earlier point in history but have been abandoned today. An example in case is the principle of determinism that was toppled by quantum mechanics. In this light, tomorrow's scientific theories may well lie beyond today's framework of ampliative criteria of scientific reasoning. The scientist nevertheless has a method of retaining a reasonably stable framework. She can assume, based on extensive and coherent historical data, that most of the principles of scientific reasoning valid today will survive the next stages of empirical testing. She thus can refer to a vague notion of a core of today's scientific principles which will remain viable at the next stages of scientific enquiry. That core of present-day scientific principles then can be deployed as a framework for the claim of limitations to scientific underdetermination.

It is important to understand that a precise specification of scientificality conditions is not necessary for carrying out an inference to limitations to underdetermination in a coherent way. The inference is based on the observation of frequent predictive success in the research field. On that basis, it is inferred first that scientists apply a framework of scientificality conditions which are quite stable, and second, that scientific underdetermination is limited within that framework. The precise nature of the framework is never made explicit in the argument.

One further consideration may alleviate doubts about basing scientific reasoning on foundations of such a shaky nature. The deepest foundations of

scientific reasoning always tend to be the ones which are the most difficult to stabilize philosophically. Just think of the problem of induction, the intractability of which stands in stark contrast to its pivotal role in all scientific reasoning. Part II of the book will sharpen the focus of this comment. It will turn out that assessments of limitations to underdetermination are by no means confined to the context of empirically unconfirmed theories. Rejecting its capability of generating scientific knowledge due to its unstable philosophical foundations would threaten much of what the scientific observer is used to take as stable scientific knowledge.

3.3 The scope of the three arguments

The three arguments of limitations to scientific underdetermination, if applicable, suggest a scarcity of possible alternatives to a given empirically unconfirmed theory that has been developed for theoretical reasons. Based on the three arguments, limitations to underdetermination are taken to be sufficiently strong for justifying the belief that the theory in question will get empirically confirmed once the critical experimental tests can be carried out. How can we understand the scope of that claim?

High energy physics knows one crucial parameter that defines the range of experimental testing: the energy provided for generating new particles in deep inelastic scattering. If a high energy theory predicts new particles at some characteristic energy scale, predictive success at that scale can be explained by a scarcity of theoretical alternatives which are predictively distinct at that scale. Meta-inductive reasoning then can lead from the observation of regular success of that kind to scarcity of alternatives with regard to the next step of empirical testing at some higher energy scale.

Note that this inference does not amount to an absolute limit to the number of empirically distinct scientific theories which fit the available data. This point can be seen most clearly in MIA, where empirical testing at a meta-level can only provide support for limitations to underdetermination regarding the immediate tests of core predictions made by the theory in question. The fact that those core predictions have been confirmed does not justify the claim that the theory will never be superseded. Attributing a low probability to the existence of alternatives which are predictively different at the next stage of empirical testing is fully consistent with the expectation that an infinite sequence of ever higher energy scales lies beyond that next empirical step. That, however, would imply that any number of alternatives could be made probable by taking into consideration a sufficiently high number of future experimental steps. The empirical

data establishes only one restriction to those new theories: they must have the known theory as their low energy effective theory (i.e. as a good approximation up to some level of accuracy) at that theory's characteristic energy scale.

Let me illustrate this point by looking at the prime example of predictive success in the high energy physics research program, the standard model of particle physics. As discussed in the context of MIA, the predictive success of the standard model is constitutive for the current trust in empirically unconfirmed theories like string theory. However, the success of the standard model did not exclude (and its creators did not intend to exclude) that more fundamental modifications of the present physical theories might also be capable of curing the problem of the renormalizability of nuclear interactions.[2] For example, instead of relying on gauge symmetries, one could have ventured already in the 1960s to make the step towards extended elementary particles, a step that was later realized by string theory. To believe in the standard model in the early 1970s merely meant assuming that any more far-reaching change of physical postulates, in as much as it would be successful, would itself imply the standard model predictions. This assumption has been vindicated by the subsequent development of physics. Extended elementary particles have emerged as a (potential) next scientific step but it turned out that their introduction, if consistently done, implies gauge theory as well.[3]

Relying on this kind of example, the physicist who applies MIA can only infer the new theory's viability at the next steps of empirical testing (irrespective of the chances for actually taking those steps in the foreseeable future). MIA does not address the question as to whether the theory to be assessed is absolutely true, whether it is empirically adequate under all possible evidence or the like. It does not imply that the theory will never have to be improved or extended by new theoretical principles or even be superseded by a fundamentally different theoretical concept. Formulated in terms of scientific underdetermination, MIA does not support the claim of an absolute limitation of the number of possible scientific theories. It merely establishes the claim that the number of theories which can be distinguished at the next steps of empirical testing is probably strongly limited. Theories which cannot be distinguished from each other at the next steps of empirical testing are not counted as different theories. We want to call a claim that asserts a limited number of possible theories which fit the available data and are

[2] In fact, in the case of the standard model technical reasons did suggest early on that the theory at some stage would have to be embedded within a wider theoretical framework. Those considerations were instrumental for developing those theories beyond the standard model which were presented in Chapter 1.

[3] String theory can only be consistently formulated in a way that makes its low energy effective theory a gauge theory (see e.g. Polchinski, 1998, Chapter 12).

empirically distinct at the next stages of empirical testing a claim of *local* limitations to scientific underdetermination.

UEA in itself does not reach out beyond the limits just described either. Unexpected explanatory interconnections pertaining to the given theory's characteristic scale may be explained by limitations to scientific underdetermination at that scale but do not exclude the possibility that new theoretical options which share the same explanatory potential might open up at higher scales.

The situation regarding NAA is a little more complex. In its most radical interpretation, NAA can be interpreted as suggesting that no experimentally viable alternative to the theory in question exists at all. In other words, NAA might be taken to suggest that the theory in question will never be superseded by any new theoretical framework because no alternative scientific theory that can reproduce the available data is possible. Such a radical claim shall henceforth be called a final theory claim. In most contexts of physical research, a final theory claim would be rendered highly implausible by the fact that the theory under investigation is not totally universal. A more advanced theory that constitutes a next theoretical step towards higher universality, however, would constitute a conceptual alternative of the kind ruled out by the final theory claim. For all theories which are not totally universal, the more modest interpretation in terms of *local* limitations to scientific underdetermination thus seems appropriate.

Now, string theory is a universal theory of all interactions. In this light, it does not seem absurd in the case of string theory to make a final theory claim. However, as argued above, the empirical corroboration of NAA must be based on MIA, which in turn can only establish local limitations to scientific underdetermination. Therefore, the significance of NAA as an argument for a final theory remains entirely unsupported within the framework discussed up to now. In order to make a serious case for a final theory, one would need other independent arguments which make plausible the additional step. We will pick up this thread of reasoning in Part III of this book where we will present two kinds of arguments which actually hint in this direction.

3.4 Non-empirical theory assessment as inference to the best explanation

We have now developed the full line of reasoning that infers a theory's chances of being viable at upcoming stages of empirical testing from assessments of scientific underdetermination. In the following two sections, we are going to relate that argument to two influential schematizations of scientific reasoning in current philosophy of science. In this section, we will define local assessments

of scientific underdetermination in terms of inference to the best explanation (IBE). The following section will analyze it from a Bayesian perspective.

The three arguments of non-empirical theory assessment which were discussed in this chapter are all examples of IBE. IBE is a form of inference to the truth or viability of statements that was first defined by Peirce and is generally taken to be constitutive of scientific reasoning. It goes beyond deductive reasoning and enumerative induction by inferring the viability or truth of statements based on the claim that they constitute the best explanation of a certain set of empirical phenomena. For the sake of simplicity we will henceforth discuss the subject in terms of truth, keeping in mind that IBE-type reasoning may also be deployed with regard to more limited scientific goals than truth. Peter Lipton (2004) and Alexander Bird (2007), who have offered recent analyses of IBE, distinguish two phases of inference to the best explanation. A first phase consists in finding a number of potential explanations from the pool of all possible explanations. A second phase then ranks the explanations found in phase one and eventually, if certain conditions are met, selects the "best" one and infers its truth.

The success of IBE obviously requires that the scientists select a set of explanations in phase one that includes the true explanation. In order to trust IBE, scientists must be confident that the true theory does not belong to the set of unconceived alternatives. Strictly speaking, IBE thus always involves an element of reasoning that amounts to assessments of scientific underdetermination. Normally, however, reconstructions of IBE do not focus on that aspect but are much more concerned with the selection process of the best theory among the selected alternatives (see Lipton, 2004, and Bird, 2007). The implicit idea seems to be that, if the best explanation is very convincing indeed, we may be justified to disregard the threat of unconceived alternatives. As an immediate consequence of this perspective, IBE is taken to be convincing only in cases where the best-known explanation is known to be very good. In particular, the best-known explanation should offer convincing explanations for all or most phenomena in the relevant domain and all or most of its implications should be known to cohere with the data. The quality of an explanation whose core predictions have not been confirmed yet cannot be assessed to a sufficient degree, which means that theories lacking empirical confirmation normally are not taken to merit trust based on IBE. IBE as normally understood thus is in agreement with the canonical understanding of theory assessment.

The present analysis is concerned with situations where theories are trusted despite a lack of confirming data. The general scheme of IBE is fully applicable in that case. However, the reasons for trust in IBE cannot be derived from the quality of the known scientific explanation alone. It must involve assessments

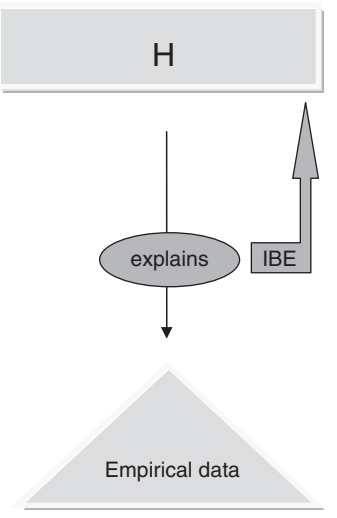

Figure 3.1 IBE at the level of scientific analysis.

of scientific underdetermination of the form discussed in the previous sections. How can this novel scenario be construed in terms of IBE? Let us try to apply IBE on the case of string theory step by step. Already at the first step, the selection of theories, the situation does not quite match conventional cases of IBE: no ensemble of theories can be formed because only one candidate theory has been found. Indeed, contexts of this kind have been explicitly addressed by Alexander Bird (2007). He speaks of "inference to the only explanation" in those cases. In the absence of any other known explanations, one must infer that the only known explanation H is the viable theory. That kind of reasoning is schematized in Figure 3.1. A triangle signifies empirical input, a rectangle a theoretical statement (in the given case the theory H_c), a simple arrow an explanation and the broad arrow an inference to the best explanation (from a theory's explanatory character to its validity).

Since theory H is not empirically confirmed in the given case, we do not know whether H would still be a good explanation once all data is in, which means that IBE seems very weak. In order to strengthen it, we have to infer the probable viability of H on different grounds: we have to assess limitations to scientific underdetermination. That assessment once again takes the form of IBE, albeit at a meta-level. We understand that statements of limitations to scientific underdetermination (let us denote them Y) can explain the three kinds of observations discussed in the previous section: (a) the observation that scientists have not found any alternatives despite looking for them; (b) the

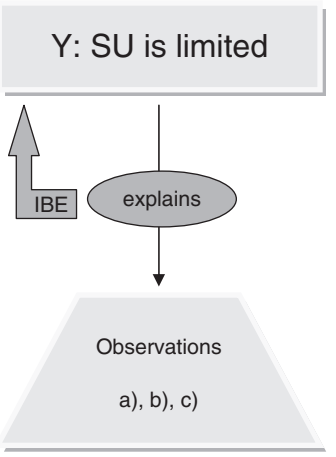

Figure 3.2 IBE at the meta-level.

observation that scientists developed the theory without considering a number of explanatory interconnections which emerged later on; and (c) the observation that in comparable cases within the research field scientific theories eventually tended to turn out predictively successful. Since we find no other equally satisfactory explanations of observations (a), (b) and (c), we infer the truth of Y. IBE at the meta-level of limitations to scientific underdetermination is represented in Figure 3.2. The trapezoid denotes the observations (a)–(c) which do not constitute empirical data in physics but are relevant for theory assessment. If sufficiently strong, the statement Y on limitations to scientific underdetermination now provides direct support for the viability of theory H and therefore assumes the role of a crucial element of IBE at the ground level. Non-empirical theory assessment thus involves IBE both at the ground level and at the meta-level, as drawn in Figure 3.3.

The presented analysis demonstrates that the concept of IBE can be extended to the case of non-empirical theory assessment. In fact, it even says a little more than that. It was mentioned in the beginning that assessments of underdetermination implicitly must enter IBE even in cases where empirical theory confirmation is available. Thus, the scheme I present may be seen as a completion for all IBE-type reasoning; to the extent that scientific reasoning is genuinely based on IBE, it must, at least implicitly, include a second level of IBE that allows for assessments of limitations to scientific underdetermination. In other words, assessments of scientific underdetermination may be more central to scientific theory assessment in general than one would assume at first sight. Part II of this book will further pursue that train of thought.

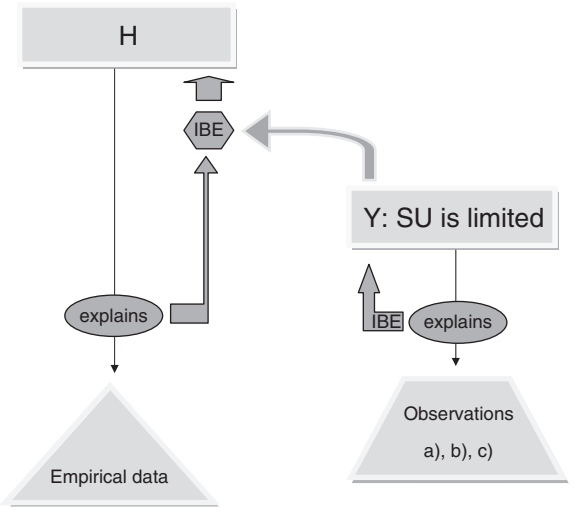

Figure 3.3 Overall inference scheme.

3.5 A Bayesian analysis

The currently most influential understanding of theory confirmation is based on Bayesian reasoning.[4] We have already encountered the Bayesian approach in Section 2.1. Theory confirmation in a Bayesian sense is provided by empirical data E that raises the probability of the truth (or, more generally, the viability) of theory H. As noted in Section 2.1, such empirical data is canonically understood to be constituted by data that is predicted (either deductively or statistically) by theory H. Therefore, it was argued, Bayesian theory confirmation can serve as an exemplification of the canonical paradigm of theory assessment. Does this imply that a Bayesian understanding of theory confirmation is at variance with non-empirical theory assessment of the kind presented in previous sections or is it possible to formalize non-empirical theory assessment along Bayesian lines of reasoning? Recent work by Dawid, Hartmann and Sprenger (in press) has analyzed this question and demonstrated that core procedures of non-empirical theory assessment do indeed constitute theory confirmation in a Bayesian sense. The following pages will not carry out a formal analysis but present the general layout of the argument.

Nothing in the Bayesian framework explicitly prescribes that data E must be predicted by theory H in order to constitute confirmation of H. The core reason

[4] Introductions to Bayesian epistemology in scientific reasoning are Bovens and Hartmann (2003) and Howson and Urbach (2006).

why one normally looks at data E that is predicted by H lies in the fact that the inequality $P(T|E) > P(T)$ that establishes confirmation (T once again being the statement that H is true or viable) can be proved straightforwardly in this case: if E follows deductively from H, then $P(E|T) = 1$. Since we cannot exclude other data than E in the absence of theory H, we have $P(E) < 1$, from which the above inequality follows based on Bayes' theorem. If E follows statistically, the situation is less rigid but similar. An extension of the Bayesian approach towards data that is not predicted by H must offer a strategy how to establish confirmation in the absence of the described simple line of reasoning. In fact, the discussion carried out in the previous sections has already provided us with the necessary tools to that end. It just remains to formalize them within the Bayesian framework.

Let us assume some evidence F of a kind that is not predicted by theory H. We call such evidence "non-empirical" with respect to H. Moreover, let us assume that F is predicted by another hypothesis Y. We thus can expect – under normal circumstances – that data F confirms hypothesis Y. Finally, let us assume that the truth of Y raises the probability of the truth of H. In that case, F raises the probability of the truth of H as well and therefore constitutes confirmation of H in a Bayesian sense.[5] In order to find a Bayesian formalization of non-empirical theory assessment, we therefore have to find a suitable candidate for hypothesis Y. The analysis of the previous sections suggests a specific candidate: a hypothesis on the number i of alternative theories that account for the available data and fulfil a given set of scientificality conditions. We introduce an infinite valued variable Y, where a value Y_i corresponds to the statement that there exist i alternative theories. Furthermore, we define Y_i^- as the statement that the number of possible alternative theories is lower than i. Y_i^+ denotes the inverse statement that there exist at least i alternatives. Simple statements which posit limitations to scientific underdetermination would be of the kind Y_i^-. More generally, a Bayesian formalization of a statement on limitations to scientific underdetermination attributes a probability $P(Y_i)$ to each Y_i.[6]

[5] This general strategy of theory confirmation has been discussed before in the context of the novel confirmation debate. See e.g. Maher (1988), Kahn, Landsberg and Stockman (1992), Barnes (2008) and Dawid (in press).

[6] There is a close relationship between assessments of limitations to scientific underdetermination and the "catch-all hypothesis" (Shimony, 1970), which is the hypothesis that we understand the spectrum of possible alternative theories sufficiently well to attribute a low value to $P(E|\neg T)$. (Low values of $P(E|\neg T)$ are required for having good confirmation by empirical evidence E.) A strong statement on limitations to scientific underdetermination can provide justification for low $P(E|\neg T)$ if one knows about the existence of any alternative scientific theory that does not predict E. Note, however, that the assumption of strong limits to scientific underdetermination is by no means the most common strategy of justifying low $P(E|\neg T)$. Low $P(E|\neg T)$ often results from simply assuming a very wide spectrum of possible alternative theories which do not predict E.

On that basis, a formalization of the argument of no alternatives (NAA) can be carried out. The relation between F and Y is most immediate. One can define a two-valued variable F_A with value $\mathrm{F_A}$ corresponding to the statement that scientists have not found any alternatives to theory H which are (expected to be) consistent with the available data and $\neg\mathrm{F_A}$ corresponding to the statement that alternatives have been found. Now, one assumes, in accordance with the arguments of Section 3.1, that it is more likely that scientists do not find any alternatives to theory H the fewer the number of possible alternative theories that exist. The weakest way to formally implement this condition is to assume $P(\mathrm{F_A}|\mathrm{Y}_i^+) \leq P(\mathrm{F_A}|\mathrm{Y}_i^-)$ for all i and $P(\mathrm{F_A}|\mathrm{Y}_i^+) < P(\mathrm{F_A}|\mathrm{Y}_i^-)$ for at least one $i > 0$. It can then be shown that

$$\langle Y\rangle := \sum\nolimits_{i=1}^{\infty} P(\mathrm{Y}_i)\mathrm{Y}_i > \langle Y\rangle_{\mathrm{F}} := \sum\nolimits_{i=1}^{\infty} P(\mathrm{Y}_i|\mathrm{F_A})\mathrm{Y}_i.$$

$\mathrm{F_A}$ thus lowers the expectation value of Y and thereby serves as an indicator for limitations to scientific underdetermination. It can also be shown that $\langle Y\rangle_{\mathrm{F}}$ is finite even if $\langle Y\rangle = \infty$ for a large class of scenarios. Second, again following the line of reasoning in Section 3.1, one introduces rather weak formalized versions of the assumptions that both the viability of theory H and the occurrence of F are the more likely the lower the number of possible alternatives. Furthermore, one assumes that T is conditionally independent of $\mathrm{F_A}$ given Y. This amounts to the assumption that the belief in the empirical viability of H is not influenced by the information that scientists did not find alternatives if the number of possible alternatives is known exactly.[7] The Bayesian network representation of the given setup (including a variable D to be explained below) is given in Figure 3.4. Under the stated conditions, one finds

$$P(\mathrm{T}|\mathrm{F_A}) \;>\; P(\mathrm{T}),$$

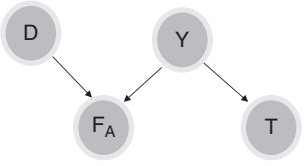

Figure 3.4 The Bayesian network representation of NAA.

[7] Strictly speaking, one would just need the very plausible assumption that $\mathrm{F_A}$ does not *increase* the probability that the unconceived alternatives are true. If $\mathrm{F_A}$ would *decrease* the probability of those theories being true, that would imply a further increase of $P(\mathrm{T}|\mathrm{F_A})$ and thus just add another element of theory confirmation by $\mathrm{F_A}$.

which means that F constitutes confirmation of theory H. In other words, under plausible assumptions NAA does amount to theory confirmation in a Bayesian sense.

The Bayesian analysis also shows the limits of NAA and elucidates the relatedness between NAA and MIA (the meta-inductive argument from predictive success in the research program). As already discussed in Section 3.1, the no alternative argument suffers from the fact that the complexity of possible solutions to the given scientific problem in conjunction with limitations to the scientists' capability might also explain the absence of known alternatives to theory H. If the complexity of the scientific problem is the true explanation of F, however, no inferences to the viability of H can be drawn. The problem is that F can never distinguish between an actual lack of alternatives and the scientists' insufficient capabilities. This threatens the significance of the no alternative arguments due to the entirely subjective character of initial probabilities. If we start with specific subjectively chosen prior probabilities for the specific statements Y_i^- and D_j^+ (where D_j parameterizes the complexity of the scientific question in comparison with the scientists' capabilities), future evidence F cannot change the ratio between these probabilities:

$$P(\mathrm{Y}_i^-)/\mathrm{P}(\mathrm{D}_j^+) = \mathrm{P}(\mathrm{Y}_i^-|\mathrm{F_A})/P(\mathrm{D}_j^+|\mathrm{F_A}).$$

Thus, even though $\mathrm{F_A}$ formally constitutes theory confirmation, the probabilities of the viability of H extracted from $\mathrm{F_A}$ do not converge under future evidence $\mathrm{F_A}$. The confirmation value of $\mathrm{F_A}$ therefore remains largely subjective. Once one believes that D_j^+ is the far more probable explanation of $\mathrm{F_A}$ than Y_i^-, no further observations $\mathrm{F_A}$ will ever constitute strong confirmation for Y_i^-. In order to remove this deadlock, we need evidence that supports Y_i^+ without supporting D_j^+.

MIA can provide such evidence. As argued in Section 3.1, predictive success within the research program can be understood as an indicator that scientific underdetermination within the research program tends to be limited. On the other hand, predictive success clearly does not suggest that scientists are not clever enough for finding viable theories in the research field. Therefore, calling data about predictive success in the research program $\mathrm{F_M}$, we have

$$P(\mathrm{Y}_i^-)/P(\mathrm{D}_j^+) > P(\mathrm{Y}_i^-|\mathrm{F_M})/P(\mathrm{D}_j^+|\mathrm{F_M}).$$

Increased probabilities $P(\mathrm{Y}_i^-|\mathrm{F_M})$ extracted from the meta-inductive argument now can be used as priors for the argument of no alternatives. The meta-inductive argument thus is capable of strengthening the significance of the no alternatives argument.

The line of reasoning sketched above establishes two important points. First, it is justified to call non-empirical theory assessment theory confirmation in a

Bayesian sense. And second, assessments of limitations to scientific underdetermination provide a workable foundation for the Bayesian formalization of non-empirical theory confirmation. The core tenets of Chapter 3 thus are supported by a Bayesian analysis. Though the Bayesian approach at first glance seems to instantiate the canonical paradigm of theory assessment, in the end it turns out to imply the possibility and potential significance of non-empirical theory assessment.