Chapter 3: From Confirmation to Explanation

Stephan Hartmann and Jan Sprenger

November 17, 2010

Abstract

This chapter motivates a Bayesian analysis of explanatory power by elaborating salient paralles to confirmation theory, and transferring arguments from Bayesian accounts of confirmation and support. Qualitative accounts of explanation and confirmation are compared in order to strengthen the structural similarity of both concepts. This comparison prepares an argument by analogy with respect to shifting the focus to a quantitative analysis. Then, we argue for a subjective, Bayesian model of explanatory power where the rationalization of the explanandum plays a central role. This account is embedded into the philosophical tradition, defended against some philosophical concerns, and critically appraised.

1 Introduction

The search for an adequate explication of scientific explanation has been going on for the last sixty years, and has remained inconclusive. Starting with Hempel and Oppenheim's (1948) famous Deductive-Nomological (D-N) model, lots of sophisticated proposals have been made, but none of them succeeded at rebutting the flood of objections and counterexamples. There is no consensus about whether and how a scientific explanation can be characterized, using a set of precise and unambiguous conditions.

In the past, the *quantitative* aspect of explanation has been neglected in favor of a charaterization in terms of necessary and sufficient conditions. But it is all but natural to ask to what extent a phenomenon is explained by some hypothesis, and how we can quantify the explanatory power of the latter. This is the project that shall be pursued in this and the following chapter.

To better place this research project in the debate on scientific explanation, a crucial distinction should be made. Explanatory relations are usually explicated in either of two distinguishable ways. First, one may be interested in

uncovering the nature or essence of the explanatory relation, the components and conditions that, when present, *thereby* make a theory-phenomenon pair an explanation. Call accounts of this first type "metaphysical." Second, one might seek a characteristic formal-logical structure that commonly attains between explanans and explanandum concomitantly to the metaphysical relation. Call accounts of this type "logical."

In keeping with the general spirit of this book, it is not surprising that we aim at a logical, and not at a metaphysical account of explanation when quantifying the explanatory power of a theory-phenomenon pair. This should not be conflated with, nor does it imply the statement that the search for metaphysical accounts is in vain: a "logic of explanation" makes no claim to explicating the *nature* of explanation. Therefore it cannot be blamed for failing to describe explanation's nature any more than a purely metaphysical account, e.g. the causal-mechanistic account, can be criticized for its informal style. The problem of some classical accounts of explanation, such as Hempel and Oppenheim's (1948) D-N model or Friedman's (1974) unification model, was precisely that they collapsed the logical-metaphysical distinction and assumed that scientific explanation was, by its very nature, a logically analyzable thing. Unsurprisingly, such positions attracted a lot of criticism that often took the form of a counterexample: A proposition was formally classified as a valid explanation because it satisfied the relevant logical conditions when everyone agreed that it was in fact a fake explanation. These straightforward arguments were supplemented with general thoughts that contextual and pragmatic elements in scientific explanations are irreducible (Van Fraassen 1980; Ruben 1990), ostensibly showing the impossiblity of logical accounts of explanation.

Such arguments need to be taken seriously when developing a formal logical schema for describing scientific explanation. Two major ways out of the dilemma have been suggested: First, we could be satisfied with a pluralist account where different characterizations of scientific explanations serve different purposes. These different senses of scientific explanation are usually unraveled by philosophers of the special sciences who investigate what counts as an explanation in physics, biology, neuroscience or economics. For example, we might find out that explanations in physics and economics work by subsumption under a general model, whereas explanations in biology and neuroscience provide causal mechanisms.

Second, we could claim that all good scientific explanations ultimately provide the *cause* of the explanandum. Then, we would presumably have to shift our focus from a conceptual analysis of explanation to an analysis of causation;

such an analysis could then be extended to an account of scientific explanation (Hitchcock 1993; Salmon 1984; Woodward 2003).

We believe, however, that we are not required to decide this notoriously difficult question and to draw either conclusion. Rightfully, the diversity of what counts as an explanations under different circumstances has been stressed in the past, and this makes it incredibly difficult to give a purely qualitative characterization in terms of a neat set of necessary and sufficient conditions. On the other hand, the pluralist research program seems to restrict itself to an enumeration and classification of the different senses of scientific explanation. While this project may be worthwhile, we are more ambitious: we would like to capture the salient quantitative aspects of explanation, without aiming at an exhaustive description of the concept itself. This will enable us to link judgment of explanatory power to open questions in theories of explanation, such as the rationality of Inference to the Best Explanation (IBE), and the differences between explanation, causation and confirmation.

A caveat should be made, though: although our program is situated in the logical and not in the metaphysical framework, we will, from time to time, have to engage in conceptual analysis in order to argue for the adequacy of our own explicative project. It would be of little use to have a strong mathematical theory that is detached from the original concept of which we wanted to deliver a philosophical analysis. But before we move on to a quantitative, Bayesian model of explanation, we will look back at history and to review a venerable qualitative model of explanation – the Deductive-Nomological (D-N) model, and compare it to qualitative models of confirmation. This analysis will motivate a Bayesian framework for modeling explanatory power.

2 Confirmation: From Successful Predictions to Inductive Arguments

The oldest attempt to formally characterize scientific explanation with elementary formal tools is Hempel and Oppenheim's (1948) D-N model – explanations are valid deductive arguments that contain at least one empirical law among their premises:

Deductive-Nomological Model of Explanation The pair $\langle H, A\rangle$ is an explanans for explanandum E if and only if

1. H is a true empirical law,[1]

[1] That the explanans must be true is, for Hempel and Oppenheim (1948, 137), rather an

2. E follows deductively from $H.A$ ($H.A \vdash E$),

3. E does not follow deductively from A alone ($A \nvdash E$)

In other words, an empirical law H provides, together with auxiliary hypotheses A, an explanation of E if and only if E can be derived from H and A, taken together, and H is crucially required for the derivation. Briefly, explananda are subsumed under covering laws. The structural equality to the most venerable account of confirmation, the Hypothetico-Deductive (H-D) model of confirmation, is evident:

Hypothetico-Deductive Model of Confirmation A piece of evidence E confirms a hypothesis H relative to background assumptions K if and only if

1. $H.K$ is consistent,

2. E follows deductively from $H.K$ ($H.K \vdash E$),

3. E does not follow deductively from K alone ($K \nvdash E$)

The idea of H-D confirmation is thus that a piece of evidence E confirms a theoretical hypothesis H if it is a prediction of H, where background knowledge K help to create the deductive link between H and E. So in both cases, the evidence is predicted by a body of theoretical propositions. Some of them have auxiliary character, others are central for the explanation/confirmation relation. Therefore, the "systematic power" of a theory (Hempel and Oppenheim 1948, 164), the ability to derive as many predictions and data as possible, is not only an asset for giving deductively-nomological explanations, but also for confirming the theory in the hypothetico-deductive sense.

Both models build, either implicitly or explicitly, on an account of lawlikeness. For the case of explanation, this requirement is motivated by observing that not all generalizations which accidentally happen to be true, increase scientific understanding and can figure in a valid explanation. The famous textbook illustration is that true sentences like "no gold spheres have masses greater than 100 tons" differ from sentences such as "no uranium spheres have masses greater than 100 tons": the latter is true because of fundamental properties of uranium, namely that it is a highly radioactive and unstable material, whereas the former is (probably) contingently true, but void of explanatory content. Similarly,

empirical than a *logical* condition of adequacy, so that we will neglect this requirement in the remainder.

"all men who take birth control pills avoid pregnancy" is a universal and true sentence, but certainly no *explanation* of why men do not give birth to children.

To distinguish explanatory from non-explanatory deductive arguments, one might therefore conclude that only empirical laws have the power to explain a phenomenon (Carnap 1966). The above statement about uranium spheres would then qualify as an empirical law, the analogous statement about gold spheres not. In a similar vein, a H-D account of confirmation without an account of lawlikeness opens the door to Goodman's (1983) new riddle of induction. Take the hypothesis "all emeralds are green" which is, apparently, confirmed by the observation of a green emerald. But now consider the hypothesis "all emeralds are grue" where "grue" is defined as green if the emerald has been examined in the past, and as blue if not yet examined. According to most natural formal accounts of confirmation, including the H-D model, the observation of a green emerald also confirms the hypothesis that all emeralds are grue.[2] The same case can be made for the analogously constructed predicates "gred", "grellow", and "grite". Thus, (infinitely many) mutually incompatible hypotheses are confirmed by one and the same piece of evidence! While this awkward consequence may be *logically* acceptable, it is *epistemologically* problematic (Fitelson 2008): if empirical confirmation is supposed to guide us to the appraisal of some theories over others, purely formal account of confirmation (such as the H-D model) cannot achieve this goal since there is an infinity of hypotheses confirmed by the total available evidence. Thus, the problem of lawlike vs. accidental generalizations equally affects the H-D model of confirmation and the D-N model of explanation. Both models share the same syntactic structure and the same problem of defining what kind of hypothesis counts as confirmable or explanatory.

Given all that, it is not surprising that both accounts resemble each other with respect to a further challenge as well: discerning relations of evidential/explanatory relevance. In both models, tacking irrelevant propositions to the confirmed hypotheses/the explanans preserves the confirmation/explanation relation. Consider the hypothesis X which is irrelevant to the hypothesis under test H which qualifies as a (possible) empirical law.[3] Now suppose that relative to background knowledge K, we can deduce E from $H.K$, but not from K alone.

[2] More exactly, we have to think of a (known) emerald whose color is examined. Then, both the hypothesis "all emeralds are grue" and "all emeralds are green" agree in their predictions: emeralds found in the past have to be green. Hence, both hypotheses are confirmed by that evidence.

[3] To pump intuitions further, we could even suggest that X may be in tension with H as long as they are logically consistent.

Then it must also hold that $H.X.K \vdash E$ so that according to the above formal schemes, E is explained by $H.X$, or it confirms $H.X$. This sounds absurd since X is completely irrelevant to the evidence (Salmon 1984; Ruben 1990). Think of, for example the hypothesis that "all gases expand whenever heated *and* frogs are green"; such composite hypotheses do not explain the behavior of a gas when it is heated. Neither do observations about the behavior of the heated gas confirm the composite hypothesis.

The structural closeness between the dominant qualitative accounts of confirmation and explanation, and the number of problems that they share, suggest that both concepts might resemble each other at the quantiative level, too. In confirmation theory, the Bayesian approach has been extremely successful both as an answer to the above challenges and as an extension of qualitative confirmation theory (we will recap some merits below). Those successes substantiates the hopes that a Bayesian account will fare equally well with respect to explanatory power. The rest of the chapter explores the prospects for analysing explanatory power in Bayesian terms.

3 Towards a Bayesian Account of Explanatory Power

Setting up a quantitative account of explanation, or more precisely, explanatory power, is a natural project. Not all explanations are equally convincing: It would be good to be able to say that hypothesis H explains evidence E better than hypothesis H' – and to which degree it does so –, instead of just saying that both count as explanations of E. So the purely qualitative accounts, like the D-N model, reach their limits. But why should be go Bayesian for measuring explanatory power? The answer is twofold: First, explanatory reasoning is a special form of reasoning under uncertainty where considerations of plausibility play a central role. Second, while causation is usually thought to focus on objective relations between events or types of events in the world, explanation has an inherently subjective dimension. In other words, we focus on the epistemic dimension of explanation, on the impact that sound explanations have on our rational degrees of belief, on the extent that a person's convictions are changed by incoming evidence. This (vaguely described) concept is called explanatory power, and Bayesianism is tailor-made for explicating it. (In the next section, we devote more space to the demarcation between explanation an causation.)

It might be argued that the Bayesian approach does not tell us anything

about scientific explanation. Analyses of explanatory power that focus on a person's subjective degrees of belief are ostensibly missing the point: they are not about explanation in science, but about how scientists change their views. The epistemic circumstances of giving explanations are arguably much less important than the essence of scientific explanations, than the way they create understanding and foster scientific progress. In other words, Bayesians describe *epiphenomena* of scientific explanation rather than explanation itself.

Notably, the very same critique has been raised with respect to Bayesian confirmation theory. The goal of qualitative accounts – to give reliable criteria for when an observational statement confirms a scientific theory – has, with the advent of modern Bayesianism, gradually been replaced by conceiving confirmation as a subjective relationship about how a person's credences about a hypothesis are affected by some evidence, by making increase in degree of belief the central notion. But that move, although familiar to modern philosophers of science, has been far from uncontentious: in crucial episodes from the history of science where prominent theories are confirmed, reconstructions in terms of degrees of beliefs seem to miss the point. Strength of confirmation is, apparently, an *objective* relation between theory and evidence, and not dependent on the prior degrees of belief of any individual scientist, or of the community as a whole. Glymour (1980) gives a couple of incisive case studies, especially from physics.

Nevertheless, the subjectivist research program in confirmation theory, that we introuced in the first chapter of the book, has been extremely successful. To briefly recap some successes: the paradox of the ravens could be reconstructed and solved by means of a Bayesian account (Fitelson and Hawthorne 2010). The tacking paradoxes are mitigated by the observation that for reasonable measures of confirmation, the irrelevant conjunction is much less confirmed than the original hypothesis (Hawthorne and Fitelson 2004). Finally, the Duhem-Quine problem lends itself to a Bayesian analysis (Earman 1992). And so on. Even more, confirmation theory has connected to the foundations of statistics and interacted with research in statistics that centers around the question about how to measure probabilistic evidence, or goodness of fit between theory and data (Good 1983; Howson and Urbach 1993; Royall 1997). Given that statistics is taking more and more ground in the empirical sciences, the rewards of these advances can hardly be overestimated. Finally, recent research has connected to with the psychology of reasoning, as witnessed by several collaborative projects between psychologists and confirmation theorists (Crupi, Tentori and Gonzalez 2007; Crupi, Fitelson and Tentori 2008).

Summing up, the power of the Bayesian approach and its interaction with statistical research have justified *post hoc* the measurement of degree of support in Bayesian terms and more than compensated for a loss of accuracy that we suffer when reconstructing some historical cases. There is certainly more to scientific confirmation than a Bayesian analysis, but the failure to give a full metaphysical account does not imply that the logical project of Bayesianism is futile.

Hence, not only struggled early Bayesian confirmation theory with similar objections that dawn on the horizon for Bayesian accounts of explanations: the benefits might be similar, too. This involves the solution of refractory puzzles an challenges as well as a better account of explanatory reasoning in science. Take statistical model comparison. If all candidate models of a certain phenomenon are very improbable and data are messy, it might be interesting to ask for an analysis of the models' explanatory power, instead of asking which model is confirmed best. The latter question might be more important for inference about unknown parameters and a the decisions which we want to take on the basis of the data, but the former might give us more guidance at understanding the data-generating process, and what kind of models we should strive to develop in the future. For instance, procedures like linear regression aim at determining the explanatory power of a specific model for the data set, whereas questions of support and confirmation are secondary. Providing a quantitative, probabilistic account of explanatory power thus closes a crucial gap at the intersection of science and philosophy, allows us to compare different cases of explanation, and, by searching for an adequate measure of explanatory power, to get deeper insight into the mechanics of how the concept of explanation works. Since the parallels – or better: dualities – between explanation and confirmation on the qualitative level are very strong, we have reason to suspect that a Bayesian analysis of explanation will have similar success. That does not render the search for a qualitative account of scientific explanation obsolete; e.g., the question whether all explanations are causal or mechanistic in essence can still be meaningfully asked.

. We suggest to focus our attention on the quantitative dimension of explanation, and the epistemic impact of explanatory considerations. This field is not covered by classical conceptual analysis, but extremely important for science.

Still, the initial question has remained open so far: why should the *epistemic* dimension of scientific explanation, the change in degree of belief, stand central to our analysis? A classical answer has been articulated by Carl G. Hempel:

> the explanatory information must provide good grounds for believing

> that X [the explanandum] did in fact occur; otherwise, that information would give us no adequate reason for saying: "That explains it – that does show why X occurred." (Hempel 1965, 368)

In a more modern formulation, Hempel's idea can be captured in the principle that the explanatory power of an hypothesis relative to some evidence gives us *reason* to believe that the explanandum is true. Imagine someone saying that "X explains Y, but given X, Y is not more expected than before". This would strike us as plainly absurd, and we would refuse to count such an X as an explanation of Y. All explanatory considerations have the normative power to reinforce the belief that the explained proposition is true, and explanatory power measures the degree to which a potential explanans *rationalizes* the explanandum. And it is hard to see how the often-quoted "scientific understanding" that good explanations provide could be different from a rationalization of the explanandum in the light of the explanans. Explanations in statistical sciences, with their use of various goodness-of-fit measures, have already learned that lesson: For example, the goodness of a linear regression is often quantified by means of the sum of the residual squares – the variation in the data that the regressors are unable to account for. In other words, the sum of residual squares quantifies the explanatory gap between our regression model and the observed data, and the degree to which the model rationalizes the phenomena.

Moreover, this approach is tailor-made to analyze the role of explanations in raising the posterior probability of a theory, or in appraising one theory over another: Sound explanations give us *reason* to believe that the explanans is true. This way of reasoning is anticipated in what Charles S. Peirce's called *abductive inference*. Here is the most precise statement that Peirce (5.189) made about this particular form of inference:

> Long before I first classed abduction as an inference it was recognized by logicians that the operation of adopting an explanatory hypothesis - which is just what abduction is - was subject to certain conditions. Namely, the hypothesis cannot be admitted, even as a hypothesis, unless it be supposed that it would account for the facts or some of them. The form of inference, therefore, is this:
>
> The surprising fact, E, is observed;
> But if H were true, E would be a matter of course;
> Hence, there is reason to suspect that H is true.

This seems to directly capture the idea of rationalizing an explanandum: Take a phenomenon which we find "surprising", which stands in need of explanation.

It is rationalized by a hypothesis H and thereby also provides reason to believe that H is true. Successful explanations confirm the explanans in the Bayesian sense. This natural consequence of a Bayesian model of explanatory reasoning explains why we like to infer to explanatory successful hypotheses, and affects all accounts of theory choice and appraisal where the explanatory power of a theory is an important criterion. Note that this property does *not* commit us to a particular theory of abductive inference, such as Inference to the Best Explanation, only to acknowledging the normative pull of good explanations.

One misunderstanding should be addressed directly: since the idea of Bayesian accounts of explanation consists in measuring explanatory relevance by means of probabilistic relevance, which is a symmetric relations, Bayesian approaches do not seem to get off the ground: the asymmetrical character of explanation, e.g. in the famous flagpole/shadow case, gets lost. On a Bayesian reading, the length of a flagpole's shadow would, presumably, always be explanatorily relevant to the height of the flagpole. Isn't that an absurd outcome and a conclusive counterexample to all Bayesian analyses?

Those who believe that there is a problem assume, however, the wrong premise that the Bayesian analysis is a bidirectional conceptual analysis of the meaning of explanatory power, in the sense of specifying a set of necessary and sufficient conditions. However, it is meant as a *unidirectional* analysis of the epistemic impacts of explanation on the degrees of belief of a rational agent. We are only committing ourselves to the fact that cases of genuine explanation have to yield a positive degree of explanatory power, but not to the contrary (e.g. when reserving the flagpole/shadow case).

Hence, this objection does not prevent us from pursuing our explicative project in a Bayesian framework. Considering how fruitful a probabilistic analysis of confirmation has proven, and how similarly the both concepts have been analyzed on a qualitative level, a Bayesian explication of explanatory power is overdue, at least for the rapidly growing number of statistical explanations in science.

4 Against a Purely Causal Account of Explanation

A natural objection to our Bayesian account that builds on parallels between confirmation and explanation contends that scientific explanations are typically *causal* explanations. This objection can be developed as follows. First, proba-

	Boys		Girls	
	Applied:Admitted	Percentage	Applied:Admitted	Percentage
Sciences	50:16	32%	50:20	40%
Humanities	70:14	20%	450:100	22,2%
Total	120:30	25%	500:120	24%

Table 1: An fictitious instance of Simpson's paradox. Although the admission rates are higher for girls than for boys in the divisions of the university, the overall admission rate is higher for boys.

bilifying explanation in terms of subjective degrees of belief apparently neglects that genuine scientific explanations proceed by giving an objective, real-world cause of the explanandum. Causality appears – at least in our pre-theoretical, intuitive understanding – to be a relation between objects or processes in the real worlds and thus independent of personal belief states that we consider to be so crucial to explanatory power. Starting from the plausible presumption that many explanations are causal, our Bayesian analysis is apparently committing a category mistake because we analyse causal, objective explanations in subjective terminology.

Such a critique is supported by ontic accounts of causality such as the mechanisms account proposed by Machamer, Lindley and Craver (2000) or Salmon's (1984) transmission mark account. But also the probability-based accounts of causality that enjoy a lot of popularity these days (Suppes 1970; Eells 1991; Hitchcock 1993; Spirtes et al. 1993) usually spell out causal relevance in terms of *objective* probabilities: "there are objective and physical relations between 'event types' that admit of probabilistic representation" (Hitchcock 1993, 338). Whether these relations need be interpreted as relative frequencies is an open question, but this understanding seems, in any case, to rule out a Bayesian interpretation.

To illustrate these worries, assume that the admission rates of the University of Neverland are way higher for boys than for girls. This suggests that there is some causal connection between gender and chance of getting admitted (maybe a gender bias in the interview, or receiving better education, etc.). Now, we split the applications into the science and humanities divisions, as displayed in table 1. Surprisingly, the picture is reversed: the admission rates are now uniformly higher for girls, although in total, girls seem to have a smaller chance to get admitted to that university. This is well-known as *Simpson's paradox*.

The explanation of the example is, of course, straightforward once you under-

stand it: girls tend to apply to more competitive disciplines (here: humanities) where there is a lower chance of getting admitted. Boys are applying relatively frequently to science programs where it is comparatively easy to get a place.

The causal link suggested by the overall figures – that being male is conducive to being admitted – is therefore refuted. But how is an explanatory analysis in terms of probabilistic relevance going to avoid the wrong conclusion that being male explains, at least partially, the higher admission rates for boys?

Following up on this point, a quantitative Bayesian analysis of explanatory power might assign some role to the prior probabilities of the explanans (why call it Bayesian otherwise?) whereas many people share the intuition that the causal relevance of a cause for an effect, or the strength of the causal link between cause and effect, should not depend on the actual distribution of candidate causes in our reference population. An example is given in Humphreys (1989) and Hitchcock (1993, 341): take a clinical trial with three randomly assigned populations that receive a moderate dose of a medical drug, a strong dose, or no dose at all. Whether the moderate dose is causally efficacious depends, on a probabilistic reading, on whether being given a moderate dose raises the probability of recovery, compared to not being given a moderate dose. That last factor, $P(R|\neg M)$, depends, however, on how many people receive a strong dose vs. no dose at all (provided that the strong dose is also effective). Most people have, at this point, the intuition that these contingent characteristics of the population cannot determine whether M is a cause of R or not. And if explanation is indeed tied to causation, then the Bayesian seems to be unable to rescue the intuition that contingent characteristics of a population should not matter for explanatory and causal relations.

Taking these observations together, our Bayesian account seems to be misguided, so much the more as it is apparently unable to make a distinction between correlation and causation. Our reply to this objection is threefold. First, and most importantly, the unidirectional nature of our Bayesian analysis does not force us to accept that all correlations are explanatory – rather, explanatory relations will be mirrored in positive correlations. Hence, there is no danger to conflate causation and correlation. On a Bayesian account, we can actually say that strong correlations are not explanatory, just because we shun a detailed conceptual analysis of what counts as a valid explanation.

Second, the examples above can also be read as arguments against an overly narrow, purely causal concept of scientific explanation. Explanation bears, by the implicit reference to someone for whom a certain state of affairs is explained, a subjective, belief-sensitive connotation. Let us push this intuition further.

Take the well-known case of the flagpole and the shadow. Usually, it is said that the length of the shadow cannot explain the height of the flagpole since clearly, the causal chain runs the other way – from the flagpole structure to the shadow that it casts. But Van Fraassen (1989) has pointed out that in certain contexts, e.g. when the shadow points at noon to a certain important spot on the ground, it is perfectly reasonable to say that the length of the shadow explains why the flagpole is ten meters high, rather then eight or twelve meters. Successful explanation is therefore not always tied to causal priority, as other everyday examples show: It is perfectly fine to say that the successful outcome of bilateral negotiations explains why one of the parties was confident to reach an agreement. And so on.

We can also give a case from "hard science". Newton's law of gravitation with its instantaneous "action at a distance" is a prime example of an explanatory highly successful covering law. Yet, it is hard to interpret causally as long as causal processes are believed to operate *locally*. Explanation has, much more than causation, a connotation of rationalizing evidence by means of positing a hypothesis, and this justifies a "subjective" understanding in terms of degrees of belief vis-à-vis an "objective" understanding in terms of causal relations. Subjectivism with respect to scientific explanation – as opposed to objectivism about causal relations – is not a threat, but rather a natural way of keeping the different functions of both concepts apart.

Coming back to Humphreys and Hitchcock's example about strong vs. moderate causation, we can now see clearly how explanation and causation fall apart. While it is counterintuitive to say that M being a cause of R depends on the size of the subpopulations in the medical trial, the overall recovery rate (that implicity refers to the composition of the population) can be *explained* by the hypothesis that already a moderate dose is effective: If there are only few people who received a strong dose, the efficacy of the moderate dose is the main explanans for high recovery rates. Whereas, if many people received a strong dose, it is not a good explanans because the efficacy of the strong dose plays a much larger part in explaining the overall recovery rate.

It might be objected that here, we engage too much in a "metaphysical" analysis of explanation rather than in the logical analysis that we promised. Yes and no. We need conceptual analysis to argue against against an overly narrow account of explanation that would rule out our proposal *a priori*, and that would miss a lot of chances to explore this exciting concept quantitatively. But the sort of analysis that we conduct does not commit us to a particular view of explanation or causation.

Third and last, there is a natural way to integrate a Bayesian analysis of explanatory power into a particular way of thinking of causation, namely into a probabilistic analysis, as proposed by Suppes (1970), Eells (1991) or Hitchcock (1993). On that account, causation amounts to probabilistic relevance in a set of relevant test situations. Usually, it is assumed that all further causes of the effect are not allowed to vary (Cartwright 1979; Hitchcock 1993); but sometimes, it is also assumed that the causes of the cause in question must be fixed.

We can then elegantly reconstruct the intuition that causal stories count as particularly good explanations. A causal connection between the potential explanans and the explanandum will very often – in particular when there is no interaction between different causes of the effect – yield strong probabilistic relevance between cause and effect. And probabilistic relevance is the common explicatum of both explanatory and (probabilistic) causal strength, although the precise way of measuring them differs. Thus, in well-behaved circumstances, a strong causal link between cause and effect will plausibly also boost the explanatory relevance of the cause for the effect. The differences between, and the interplay of causation and explanation are not problematic, but exciting!

Therefore, it remains an interesting open question how measures of causal strength (Eells 1991; Fitelson and Hitchcock 2010) can be related to the starting debate about measures of explanatory power (McGrew 2003; Schupbach and Sprenger 2010). This project will be undertaken at the end of the following chapter. From the above considerations, we have reason to suspect that they will be different since clear conceptual differences have been revealed, and these differences will likely be found back on the level of quantitative measures.[4]

[4]It might be questioned whether our Bayesian approach has substantial advantages over, say, Hempel's (Hempel 1965) inductive-statistical (I-S) or Salmon's (1971/1984) statistical-relevance (S-R) approach to explanatory power. Hempel's quantification of the strength of an I-S argument (Hempel 1965) seems to anticipate a Bayesian measurement of explanatory power, and Salmon, on the other hand, takes probabilistic (=statistical) relevance as the main idea of his proposal, just as modern Bayesians do. – There is, however, a crucial difference: both Hempel and Salmon lean on relative frequencies in relevant reference classes as the measures of explanatory strength. In the case of Salmon, this is motivated by the desire to connect an account of statistical explanations to relative frequencies in nature, whereas Hempel and his co-author Oppenheim aim at connecting their account of explanatory power to the Carnapian framework of logical probability. Thus, they are not giving a *subjective* Bayesian account, as this book proposes.

5 Conclusions

The project of explicating explanation qualitatively has suffered from many setbacks, caused by a continued strain of objections. To resolve this deadlock, we have proposed to shift attention to the quantitative, epistemic dimension of scientific explanations, in the same way that it has been done for confirmation. A Bayesian approach is the natural way of doing so. This move offers a much better and fine-grained understanding of the grammar of explanation, and has the ability to connect to the growing statistical literature on measures of confirmation and goodness-of-fit.

To motivate the project further, we argued at length for the structural duality of explanation and confirmation, where Bayesianism has been a very fruitful research program. This duality can be sustained and qualified when passing from a qualitative to a quantitative, Bayesian framework. Shifting the focus to the epistemic dimension of successful explanations, to the rationalization of the explanandum by the explanans, the chapter does not only illuminate the relationship between confirmation and explanation, but also makes a contribution to better understanding explanatory reasoning in science, and its role in theory choice and appraisal. A subjectivist, Bayesian analysis of explanatory power in terms of rational degrees of belief is not far-fetched and futile, but fruitful, feasible, and reveals exciting structural connections between causation, explanation and confirmation. The next chapter presents the details of such an account and defends a particular measure of explanatory power as a the unique feasible explicatum. Finally, we will turn to the empirical validations of our results, their implications for the theory of Inference to the Best Explanation (IBE), and a comparison between measures of explanatory can causal power.

References

Carnap, Rudolf (1950): *Logical Foundations of Probability*. Chicago: University of Chicago Press.

Carnap, Rudolf (1966): "Explanation in Science and History", in: R. G. Colodny (ed.), *Frontiers of Science and Philosophy*, Pittsburgh: University of Pittsburgh Press.

Cartwright, Nancy (1979): "Causal Laws and Effective Strategies", *Noûs* 13, 419–437.

Crupi, Vincenzo, Katya Tentori, and Michel Gonzalez (2007): "On Bayesian Measures of Evidential Support: Theoretical and Empirical Issues", *Philosophy of Science* 74, 229–252.

Crupi, Vincenzo, Branden Fitelson, and Katya Tentori (2008): "Probability, Confirmation, and the Conjunction Fallacy", *Thinking and Reasoning* 14, 182-199.

Earman, John (1992): *Bayes or Bust?*. Cambridge/MA: The MIT Press.

Eells, Ellery (1991): *Probabilistic Causality*. Cambridge: Cambridge University Press.

Fitelson, Branden (2001): *Studies in Bayesian Confirmation Theory*. Ph.D. thesis, University of Wisconsin-Madison. Available at `www.fitelson.org`.

Fitelson, Branden (2008): "Goodman's 'New Riddle'", *Journal of Philosophical Logic* 37, 613-643.

Fitelson, Branden, and James Hawthorne (2010): "How Bayesian Confirmation Theory Handles the Paradox of the Ravens", in: Ellery Eells and James Fetzer (ed.), *Probability in Science*, Chicago: Open Court.

Fitelson, Branden, and Christopher Hitchcock (2010): "Probabilistic Measures of Causal Strength", in: F. Russo and J. Williamson (ed.), *Causality in the Sciences*, Oxford: Oxford University Press. Available at `www.fitelson.org`.

Glymour, Clark (1980): *Theory and Evidence*. Princeton: Princeton University Press.

Good, I. J. (1983): *Good Thinking: The Foundations of Probability and Applications*. Minneapolis: The University of Minnesota Press.

Goodman, Nelson (1983): *Fact, Fiction and Forecast*. Fourth Edition. Harvard: Harvard University Press.

Hawthorne, James and Branden Fitelson (2004): "Re-solving Irrelevant Conjunction with Probabilistic Independence", *Philosophy of Science* 71, 505-514.

Hempel, Carl G. (1945): "Studies in the Logic of Confirmation", originally appeared in *Mind*, reprinted in: *Aspects of Scientific Explanation*, 3-46. New York: Free Press.

Hempel, Carl G. (1965): *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*. New York: Free Press.

Hempel, Carl G., and Paul Oppenheim (1948): "Studies in the Logic of Explanation", *Philosophy of Science* 15, 135-175.

Hitchcock, Christopher (1993): "A Generalized Probabilistic Theory of Causal Relevance", *Synthese* 97, 335–364.

Howson, Colin and Peter Urbach (1993): *Scientific Reasoning: The Bayesian Approach*. Second Edition. La Salle: Open Court.

Humphreys, Paul (1989): *The Chances of Explanation: Causal Explanations in the Social, Medical, and Physical Sciences*. Princeton: Princeton University Press.

Machamer, Peter, Lindley Darden, and Carl F. Craver (2000): "Thinking about Mechanisms", *Philosophy of Science* 67, 1-25.

McGrew, Timothy (2003): "Confirmation, Heuristics, and Explanatory Reasoning", *British Journal for the Philosophy of Science* 54, 553-567.

Royall, Richard (1997): *Statistical Evidence: A Likelihood Paradigm*. London: Chapman & Hall.

Ruben, David-Hillel (1990): *Explaining Explanation*. London: Routledge.

Salmon, Wesley (1971/1984): "Statistical Explanation", reprinted in Salmon (1984), 29-87.

Salmon, Wesley (1984): *Scientific Explanation and the Causal Structure of the World*. Princeton: Princeton University Press.

Schupbach, Jonah, and Jan Sprenger (2010): "The Logic of Explanatory Power", forthcoming in *Philosophy of Science*.

Spirtes, Peter, Clark Glymour, and Richard Scheines (1993): *Causation, Prediction and Search*. New York: Springer.

Suppes, Patrick (1970): *A Probabilistic Theory of Causality*. Amsterdam: North-Holland.

Van Fraassen, Bas (1980): *The Scientific Image*. Oxford: Clarendon Press.

Van Fraassen, Bas (1989): *Laws and Symmetry*. Oxford: Clarendon Press.

Woodward, James (2003): *Making Things Happen: A Theory of Causal Explanation*. Oxford: Oxford University Press.