Published as Petri Ylikoski & Jaakko Kuorikoski 2010: Dissecting Explanatory Power, Philosophical Studies 148: 201–219.

Dissecting Explanatory Power

Petri Ylikoski & Jaakko Kuorikoski

University of Helsinki

Abstract

Comparisons of rival explanations or theories often involve vague appeals to explanatory power. In this paper, we dissect this metaphor by distinguishing between different dimensions of the goodness of an explanation: non-sensitivity, cognitive salience, precision, detail, factual accuracy and degree of integration. These dimensions are partially independent and often come into conflict. Our main contribution is to go beyond simple stipulation or description by explicating why these factors are taken to be explanatory virtues in the first place. We accomplish this by using the contrastive-counterfactual approach to explanation and the view of understanding as an inferential ability. By combining these perspectives, we show how the explanatory power of an explanation in a given dimension can be assessed by showing the range of answers it provides to what-if-things-had-been-different questions and the theoretical and pragmatic importance of these questions. We also show how our account explains intuitions linking explanation to unification or to the exhibition of a mechanism.

Keywords: explanation; understanding; explanatory power; inference to the best explanation

1. Opening up the metaphor

People evaluate and compare explanations all the time. These judgments of explanatory merit are often expressed using metaphorical notions of explanatory depth and explanatory power. Philosophers too use these notions to argue for their favorite views or theories. However, the

existing philosophy of science literature does not offer much insight into these notions.[1] The situation would be acceptable if there were a standard way of applying these concepts, but this is not the case – scientists and philosophers seem to attribute explanatory power based on quite different implicit principles or without any principles at all. In this paper we endeavor to improve this situation in two ways. First, we will distinguish five dimensions of explanatory power and discuss their relationship to each other. The idea is to describe those properties of explanations that usually prompt people to attribute the metaphorical quality of ”power” to them. Second, we will present arguments that attempt to explain why improvement in these dimensions can legitimately be regarded as improvement in explanatory understanding. We will also spell out how judgments of explanatory power can involve confusion about the explanatory and evidential virtues of a proposed explanation.

We will talk about ”explanatory power”, but the points will also apply to the notion of ”depth of explanation”. There is no standardized way of distinguishing between these notions. One person’s ”depth” might be another person’s ”power” and vice versa. The only difference seems to be that ”power” is only employed in contexts in which two explanations are *compared* with respect to their explanatory qualities, whereas “depth” is also used in non-comparative contexts. For example, one might talk about increased ”depth” in situations where an explanation explains a presupposition of another explanation (for example, by providing a mechanism). We will not discuss these interesting cases in this paper.

It is important to distinguish between explanatory and evidential virtues of explanations. In practice, both of these virtues are evaluated together, but there is a conceptual difference and the possibility of confusion. One might mistake explanatory virtues for evidential virtues or vice versa. When one explanation is better supported by evidence than another, then it is doing better in evidential terms. We wish our explanations to be true, and true explanations are better

[1] Important exceptions are Miller 1986, Morton 2002, and Hitchcock and Woodward 2003.

than false ones. However, explanatory virtues are not about the likeliness of the explanatory hypothesis: they are about how good the explanation is, if it is true. When evaluating how good an explanation is, we assume that the facts it cites are true, and we evaluate how well it satisfies other desiderata we establish for explanations. This paper will focus on these desiderata.

A popular approach to scientific inference, the inference to the best explanation (IBE), assumes that there is a connection between loveliness and likeliness (e.g., Lipton 2004). The arguments presented in this paper make no assumptions about such a connection. Our question is what makes one explanation better than another when both are assumed to be true. In contrast, the proponents of IBE wonder whether explanatory merits are a reason to infer that the other hypothesis is more likely to be true. These two issues are separate, and our question makes sense even if one does not believe in the IBE. (And with some qualifications, we do not.) Of course, our account is highly relevant to the IBE. The advocates of the IBE have said surprisingly little about the explanatory virtues,[2] and our focus is precisely on the crucial notion of explanatory goodness.

Our hope is that the articulation of implicit criteria for explanatory assessment would be helpful in making sense of controversies about scientific explanation. We have in mind both controversies within the theory of explanation and controversies within the sciences concerning the merits of specific explanations. Within the theory of explanation, an account of explanatory virtues might be helpful in explaining why people find ideas such as mechanism or unification so appealing. The ability to distinguish among the different dimensions of explanatory goodness could also serve as a strong argument for the fruitfulness of the account of explanation that makes such elaboration possible. But most fundamentally, articulating the criteria for explanatory goodness is the kind of thing that a theory of explanation should be

[2] For example, Peter Lipton says virtually nothing about the criteria that make one explanation better than another (Lipton 2004: 122, 138-140).

about: what makes explanations explanatory. Furthermore, theory of explanation should not be a self-sufficient specialty in the philosophy of science. It should have a connection to real scientific practices, and its validity should be judged by its relevance to sorting out explanation-related controversies in the sciences. The validity of theories should be decided by their fruitfulness in improving scientific practices (judged by the scientists themselves), not by appeals to the intuitions of philosophers. Intuitions are not evidence for philosophical theories – they are things to be explained.

The objectives of this paper are more modest than these ultimate hopes. We will describe the five dimensions of explanatory virtue and the theoretical ideas that are needed to make sense of them. Owing to limitations of space, our discussion of individual dimensions will be brief and lacking in detailed examples, but we attempt to convey the principal ideas of our approach, which will be developed further in other publications.

2. Making sense of explanations

Our preferred approach to explanation is the contrastive-counterfactual theory of explanation (Woodward 2003). In this paper, this theory serves two roles. First, it is used to describe the different dimensions of explanatory virtue. Second, it is used to explicate why and to what extent these explanatory virtues are virtues in the first place. These roles are in principle separate: a supporter of a competing theory of explanation might accept our taxonomy of explanatory virtues but prefer a different account of their source. However, it is an open question whether competing accounts of explanation (e.g., DN-account, unification account, causal process account, pragmatic account, etc.) are up to the challenge. In these circumstances, a strong argument for a theory of explanation is that the theory can be used to

articulate and to make sense of the different dimensions of explanatory power. In the following paragraphs, we will briefly describe the central elements of this theory.[3]

The starting point of the contrastive-counterfactual approach is realistic: explanations attempt to trace objective relations of dependence. In this paper we will only discuss causal dependence, but the same basic ideas can also be applied to other kinds of dependence, such as the relation of constitution between the whole and its parts and their organization. Dependencies are objective in the sense that they are independent of our ways of perceiving, conceptualizing and theorizing about them. The kind of dependence we are interested in is modal: explanation is not about subsumption under empirical regularities, but about counterfactual dependence.[4] *X* explains why *Y* if *Y* depends on *X* in the sense that if *X* had not happened, then *Y* would not have happened either. It is commonplace to call these dependencies ”counterfactual dependencies”, although it is important to note that there is nothing counter to facts in the relation of dependence itself.

While the causal dependencies in the world are independent of our conceptualization, explanation is an epistemic activity: *an explanation can only relate things described or conceptualized in a certain way.* In other words, we always explain specific aspects of events or phenomena, not these events or phenomena themselves or as a whole. Explanations are answers to questions in the form: *why fact rather than foil*, where the foil is an exclusive alternative to the fact.[5] We want to know why things are one way rather than some other way. Here we will adopt the convention of expressing the contrast in the following manner: fact [foil], which should be read as fact rather than foil. The number of foils is not limited; there is

[3] The most detailed version of the approach can be found in Woodward 2003; see also Garfinkel 1981, Lipton 2004 and Ylikoski 2007.

[4] We follow the somewhat unfortunate but nevertheless established practice of using ”counterfactuals” to refer to all kinds of subjunctive conditionals, regardless of whether the antecedent is true in the actual world or not.

[5] Some authors have argued that fact and foil can be compatible. This claim is based on confusion: the examples cited by these authors are explanations of differences. When the structure of these explanations is explicated, it becomes obvious that we are not explaining why *x* is *A* and *y* is *B*, but why *x* is *A* rather than being *B* like *y* (Ylikoski 2007).

often more than one foil. The facts to be explained can belong to different ontological categories: they can be properties, events, quantities or qualities. In this paper we will speak in terms of variables. This terminology does not commit us to any specific ontology, but allows us to make our points more generally. The talk of variables should not be read as a preference for quantitative notions: the variables can also be qualitative properties. All that is required is that the values of the variable be exclusive alternatives to one another: when a variable has one value, it is not possible for it to have another.[6]

The contrastive idea has often been interpreted as a thesis about the pragmatics of explanation and as a claim about what people have in mind when they put forward an explanation-seeking question. These are not the most productive ways to use the contrastive idea. The idea of contrastive *explanandum* helps to make the explanation more explicit in an analytically fruitful manner (Ylikoski 2007). Spelling out the contrastive structure forces one to articulate what is the object of the explanation and in which respects we think the object could have been different. Are we explaining a singular event or a more general phenomenon or regularity? Are we addressing properties of individuals or of populations? What is the appropriate level of description: are we after micro-level details or patterns found at the macro-level? Although it is not always apparent from the surface appearance of explanations, all explanation-seeking why-questions can be analyzed and further explicated by writing down the implicit contrastive structure of the *explanandum*. This makes both the aims of the explanation clearer and the evaluation of the explanations easier. When the idea of contrastive *explanandum* is combined with the idea of the counterfactual relevance of the *explanans*, we can make judgments about what a given explanation actually explains. This can be done independently of the original explanation-seeking question.

[6] In many cases the variable can be regarded as a determinable and its value as its determinant. This idea makes it clear that the fact and its foils are in the same space of alternatives.

Quite often the original question, when articulated in contrastive terms, turns out to be a whole set of related contrastive questions. This is a good thing: smaller questions are something that we can actually hope to answer by means of empirical enquiry. Contrastive articulation is also useful in controversies over apparently conflicting explanations. Quite often the competing explanations turn out to be addressing complementary or completely independent questions. This is as it should be: we can be pluralists about explanation-seeking questions, but whether the answers are correct is still an objective matter.

In our account, the *explanans* also has a contrastive structure. In effect, we are claiming that explanation is doubly contrastive (cf. Schaffer 2005). Explanations track relations of dependence between values of variables: a change in the *explanans* variable explains the change in the *explanandum* variable. This restricts the range of possible contrasts that a sensible *explanandum* or *explanans* may have. For example, in the case of causation, the relevant changes are changes in the *explanans* variable brought about by a special kind of causal process called intervention (Woodward 2003).

The goal of explanation is the creation of understanding. As Wittgenstein pointed out, understanding should not be conceived as a special mental state, but a publicly-attributed behavioral concept akin to an ability (Wittgenstein 1953 [1997], §§ 143-159, 179-184, 321-324; Baker and Hacker 2005: 357-385). In our view, the fundamental criterion according to which understanding is attributed is the ability to make inferences to counterfactual situations, the ability to answer contrastive *what-if-things-had-been-different questions* (*what if* - questions) relating possible values of the *explanans* variables to possible values of the *explanandum* variable. This idea ties together theoretical and practical knowledge: they are not completely different notions. For example, in the case of causal explanation, explanatory understanding is crucial to our pragmatic interests, since answers to w-questions concerning the effects of possible interventions enable us to predict the effects of manipulation. Whereas

the DN-model and the associated epistemic conception of explanation conceive the possessor of understanding as a passive observer of external events, the contrastive counterfactual theory links our theoretical practices to our roles as active, goal-oriented agents (Woodward 2003). The degree of understanding conveyed by an explanation can now be defined as the number and importance of counterfactual inferences that the explanatory information makes possible. (Ylikoski 2009.)

Finally, a remark about understanding and information is in order. In an obvious sense, a good explanation is one that is informative relative to the current body of background knowledge. However, our measure of explanatory power is concerned only with how many *what if* -questions can be answered on the basis of the explanatory information (given a body of background knowledge), and not how many of these inferences are actually novel. Therefore, explanatory power is relative to the body of background knowledge, but not to its dynamics. Many powerful explanations are not perceived as such because they are simply obvious, and many shallow explanations seem interesting just because they rely on information that is somehow novel or surprising.

3. Attributing explanatory power

Usually attributions of ”explanatory power” arise in the context of (at least apparently) competing explanations. The explanations are regarded as alternatives to each other,[7] and the more ”powerful” explanation is regarded as the better explanation. Often this comparison is used as an argument in favor of the theoretical perspective from which the preferred

[7] It is important to see that two explanations do not have to be exclusive alternatives to each other. In cases of causal overdetermination, both explanations could be true and explanatory, yet it could still be said that one exemplifies explanatory virtues better than another. However, in such cases the comparison is somewhat academic: there is no legitimate epistemic reason to choose one explanation over the other. It is much better to have a more complete explanation that includes both, together with an account of their relationship.

explanation is derived. In order for this kind of comparison to be possible, the explanatory tasks should be the same or at least sufficiently similar. If the explanatory tasks were different, there would not be much point in pitting the two explanations against each other.

However, competition is not the only context in which it makes sense to compare explanations. Explanations of different things can also be compared if it is thought that they share (or should share) similar standards of explanation. This is the case when one science (or research field) is taken as an (ideal) example for or in contrast to another. In these cases, explanations in the exemplary field meet the assumed standards better. The explanations are not competing with each other; in fact, they can address very different *explananda*. The crucial thing is that the exemplar sets the standards for the field that is considered to lag in explanatory excellence. The exemplar suggests what an ideal explanation would look like.

This second context of comparison draws its importance from the fact that there are no explicit general principles for goodness of explanations. The standards of good explanation are learned (and interpreted) via paradigmatic examples (exemplars) of good explanations. The changes in explanatory standards usually take place through changes in the exemplary explanations, not via explicit discussion of the principles that govern the attribution of explanatory excellence. Similarly, the ideals of explanation travel from one discipline to another via exemplars rather than as explicit principles.[8]

”Explanatory power” can be an attribute of both theories as well as of individual explanations. We regard the latter to be the more interesting notion. The attribution of ’power’ to a theory can either refer to the theory as a template for explanations or to a filled-in version of such a

[8] The distinction between competing explanations and competing standards is useful when considering cases of interdisciplinary exchanges like disciplinary imperialism. A discipline can be regarded as a threat to another (and called imperialistic) if a) it aims to take over the *explananda* of the other; b) if it aims to change the standards of explanation by changing the set of paradigmatic examples of good explanations; or c) it does both of these at the same time. The distinction is important since quite often the criticism of these ambitions is directed at specific explanations, whereas the presenters of these explanations often regard them merely as suggestive examples of the new improved standards.

template. In the latter case we have an individual explanation, and we are talking about virtues of that explanation. When a theory provides a template for explanation, it provides a reusable sketch that can serve in constructing explanations and in the search for possible explanatory factors. Comparing such templates to actual instances of explanation would be foolish. However, the templates can be compared with respect to their explanatory power.[9]

Theories as explanation templates can be "powerful" in two different ways. First, a theory might be powerful because it can be used to generate *many* explanations. This can be understood in two ways: either to mean that the theory provides an explanation for many individual facts or to mean that the theory provides an explanation for many kinds of facts. Most people agree that the latter notion is more interesting: the actual number of token *explananda* is a contingent matter and does not make the theory in any way epistemically superior. However, it is an open question whether the latter notion is an epistemic or a pragmatic virtue. We think that it is a pragmatic virtue, but this stance is of no consequence here: the point is that this is one property that is used as a basis for attributing "explanatory power." There are some obvious problems in counting kinds of *explananda*, but these problems are not our concern: in this paper we are not interested in this quantitative notion of explanatory power, but in a qualitative notion. According to this second idea, a theory is powerful if it provides *better* explanations. Here the crucial issue is how the goodness of explanations is evaluated, which brings us back to the evaluation of individual explanations. For this reason, we will be concerned with the varying ways in which one individual explanation can be better or more powerful than another.

[9] Theories are often evaluated in terms of their future promise, not their current achievements. However, from the conceptual point of view, the achievement is more fundamental: after all, the promise is about future achievements. An evaluation of future achievements is difficult, leaving room for wishful thinking and other biases. This illusion of explanatory depth can be a very significant feature of explanation-related controversies. (Ylikoski 2009.)

4. The dimensions of explanatory power

The central claim of this paper is that there is more than one dimension of explanatory power. More specifically, we argue that there are five dimensions: non-sensitivity, precision, factual accuracy, degree of integration and cognitive salience. Some additional properties often associated with goodness of explanation, such as mechanistic detail and unification, can be accounted for on the basis of these basic dimensions, although in qualified forms.

Our second central claim is that these dimensions are separate and do not normally go hand-in-hand: improvement in one dimension does not automatically mean improvement in the other dimensions. In fact, we argue that some of these dimensions of explanatory power are systematically in conflict: there are important trade-offs between the different dimensions.

The dimensions of explanatory virtue that are judged to be more important than others always depend on the epistemic aims the explanatory information is intended to support. For this reason, the evaluation of explanations always has a pragmatic dimension. There are also some important differences between scientific disciplines and research traditions with respect to the emphasis they put on different dimensions. An improved understanding of these dimensions will help us to understand explanation-related controversies in the sciences.

The presence of these pragmatic and contextual factors does not make the assessment of explanatory power a completely subjective matter. In our view, the aim of science is explanatory understanding, which can be attributed to a cognitive agent according to the agent's inferential performance. The power of an explanation can be measured by the degree of understanding it can create. Thus, in a given dimension, the power of an explanation is more or less an objective matter. It is measurable by the range of inferences to counterfactual situations the explanatory information makes possible and by the ease with which these inferences can be made with respect to some body of background knowledge and cognitive capacities. Therefore, although our account is first and foremost descriptive, we aim to describe those properties of

explanations that give rise to attributions of ”explanatory power” by scientists – we also try to articulate why these dimensions should be regarded as proper criteria for explanatory evaluation.

4.1 Non-sensitivity

Our first dimension of explanatory power is sensitivity of the explanatory relationship with respect to changes in background conditions. The basic idea is that the more sensitive an explanation is to changes in background factors, the less powerful it is.[10] An increase in sensitivity makes the explanatory relationship more fragile, whereas a decrease in sensitivity makes it more robust. The more insensitive the explanatory generalization is with respect to the background conditions, the more independent it is from those conditions. This means that the same answer would be correct for a larger group of *what if* -questions. Sensitivity is a causal notion that refers to the context-sensitivity of the explanatory relationship and to the degree the relationship is susceptible to local causal interference. Sensitivity is not an epistemic notion that refers to the sensitiveness of our inferences with respect to changes in our background assumptions. The intuition that good explanations make their *explananda* necessary or at least less contingent can largely be accounted for by considerations of sensitivity.

Sensitivity has two forms. An explanatory relationship is less sensitive if it would continue to hold under a larger set of interventions on values of variables that are not incorporated into it. Less sensitive explanations are more powerful in that they enable inferences to more counterfactual situations in which omitted variables take non-actual values (Woodward 2006). Another form of sensitivity is the range of values that the *explanans* variable*s* can take without breaking the explanatory relationship. Sensitivity is therefore directly related to the reliability of the explanatory information; sensitive explanations provide information that is unreliable in

[10] Adam Morton equates non-sensitivity with causal depth (Morton 2002, 85-87).

situations in which there are changes in factors that are not explicitly accounted for or when the case is extrapolated to unforeseen extremes.[11] As a causal feature, sensitivity is a modal notion. An empirical generalization can be exceptionless, but the associated explanatory dependency can be extremely sensitive and *vice versa*. For example, selection or intentional planning can ensure that the background conditions required for a fragile causal dependency are always present when the relevant causal factors are. Thus, the regularity between the factors seems robust even though the dependency producing the regularity might be very sensitive.

Only rough comparisons of sensitivity can be made at a gross level by comparing overall sensitivity. This is sufficient in some circumstances: it can be used to characterize *hyper-sensitive* explanations. An explanation is hyper-sensitive if any change in the background conditions would break the dependence between the *explanans* and the *explanandum*. It is clear that such highly sensitive explanations are of value only in exceptional cases. Usually such contexts are not primarily about explanation, but about attribution of blame or diagnosing malfunctioning devices. Beyond hyper-sensitivity, there are no general principles for comparing the sensitivity of explanations. It is conceptually possible that two explanations differ only with respect to one omitted variable, and for such cases we can define the notion of a strictly less sensitive explanation[12]. In realistic cases, however, we are comparing multiple differences with respect to different background conditions. This binds the assessment of sensitivity to our epistemic goals and to the pragmatic context of explanatory evaluation in a manner that makes the idea of overall sensitivity more or less irrelevant.

The reason for the irrelevance of overall sensitivity is that not all background conditions are equally important. If all changes in background conditions were equally important, then one

[11] Note that this reliability is distinct from the epistemic evaluation of how reliably the explanatory information can be expected to be true in the first place.

[12] Explanation *A* is *strictly less sensitive* than explanation *B* if and only if there is at least one change in background conditions with respect to which *A* is less sensitive than *B* and there are no background conditions in which *B* is less sensitive than *A*.

could compute the gross sensitivity by simply counting the cases favoring each explanation. However, there is no objective measure for how large a given disturbance in a given background condition is and therefore no single correct way of aggregating them. Instead, judgments concerning the importance of changes in the background conditions are always relative to the theoretical and pragmatic context. For this reason, many disagreements about the relative sensitivity of competing explanations are not disagreements about causal facts, but about the probability or significance of possible changes in given background conditions.

As an example of the kinds of sensitivity considerations that can be weighed against each other, consider different kinds of equilibrium explanations (Kuorikoski 2007). Non-sensitivity or stability of the equilibrium with respect to the background conditions is usually taken to be an important *desideratum* for equilibrium models and the explanations derived from them. However, a crucial distinction has to be made between dynamic and structural stability. The former refers to the non-sensitivity of the equilibrium-state with respect to the initial conditions or changes in the values of the *variables* in the model; the latter refers to the non-sensitivity of qualitative properties of the equilibrium with respect to changes in the structural properties or *parameters* of the model itself. Consider an economic equilibrium explanation of a market outcome of why price p [p']? A model assuming perfect competition might be completely insensitive to historical details concerning the initial distribution of goods and thus yield an explanation for p [p'] which is maximally insensitive with respect to the initial values of variables. In contrast, an alternative model in which the market participants have varying degrees of market power might yield the equilibrium value p from only a specific set of initial endowments, but it might do this across a number of different dynamics depending on the structure of the competition. If in the actual case the market participants had only limited market power and exit from and entry to the market was relatively free, then which explanation is judged to be better depends on which kind of stability is thought to be more important or

theoretically interesting. Usually, the sensitivity that matters is structural (in)stability because the explanatory relationship of interest is the dependency between properties of the constituent parts and the qualitative properties of the equilibrium (i.e., dispositional properties of the whole) rather than the dependency between the initial state and the end state. However, since equilibrium explanations are usually quite speculative, structural stability is often more of an epistemic requirement: structurally stable equilibrium models can get away with more dubious assumptions concerning the properties of the parts.

4.2. Precision

Sensitivity is an attribute of the explanatory relationship. Our next dimension, precision, is an attribute of the *explanandum*. The question is how precisely the explanation characterizes the *explanandum* phenomenon. The more detailed the account given of the *explanandum*, the better the explanation.

Our notion of precision does not refer to factual accuracy or truth (which is already presupposed), but to the *range of the contrast space relative to the explanandum*. Precision is an attribute of the thing being explained. Let us assume that we are explaining why a person chose a certain color for a car. Compare first these two *explananda*:

[1] navy blue [any other color]

[2] blue [any other color]

It is intuitively clear that [1] is more precise than [2] and thus preferable, since the contrast space is exactly the same. In our terminology, [1] is *a sharper explanandum* than [2]. Let us next compare [2] with

[3] blue [red]

Now we have made a change in the contrast class. Here [2] is clearly better since it has *a broader range of contrasts*. The explanation of [3] could be derived from the explanation of [2] and is therefore less informative. [1] is more precise compared to [3] in both of the two ways: [1] has a sharper *explanandum* value and a broader range of contrasts.

However, things are not always so straightforward. Let us compare [2] with

[4] navy blue [other shades of blue]

Here [2] has a broader range of contrasts, whereas [4] has a sharper *explanandum*. Unequivocal comparison is not possible in this case. However, in special cases like this, it is possible to regard the explanations as being fully complementary: if we combine them, we get an explanation that is more precise on both counts. This is possible because the *explanandum* fact of [4] is a determinant of the *explanandum* of [2] and the range of contrasts of [4] is a subset of range of contrasts of [2]. In cases where these conditions are not satisfied, the combination of two explanations is not as straightforward. In such cases the comparison is based on one's epistemic aims: one has to decide which of the aspects of precision is more important. We are sometimes interested in minute differences between fine-grained details and sometimes in broader differences between types of properties or phenomena, but we are always better off with information that enables us to do more rather than less.

There is an interesting trade-off between precision and non-sensitivity. Although sensitivity of an explanatory dependency is a non-epistemic causal notion, the sensitivity of an explanation is usually increased when precision of the *explanandum* is increased. This is simply because smaller causal deviations are needed to disrupt the dependency between the *explanans* and a fine-grained *explanandum* than coarser-grained ones. Consider a comparison between an equilibrium explanation and a historical narrative of a market price p at some given time. Although a detailed historical narrative of a sequence of trades (including the psychological states of the traders) could in principle have the explanatory resources to account for a very

sharp *explanandum* (market price p [p']), this explanation would be very sensitive to counterfactual changes in what actually happened. An equilibrium explanation might reasonably explain only the fact that the price was in the neighborhood of a theoretical equilibrium point, but it would explain this in a very insensitive manner. During the process of explanatory refinement, a suitable balance between precision and sensitivity is usually found, and this balance largely determines the appropriate "level" of the explanation. For example, the historical narrative of individual trades would be an individual-level explanation of the market price, whereas the equilibrium explanation would essentially be a system-level explanation.

4.3. Factual accuracy

The consensus in the theory of explanation seems to be that explanation is factive and that false explanations are only apparent explanations. However, it is not immediately clear what the rationale of this requirement is, and most arguments for this requirement seem little more than ideological exclamations: science simply *should* aim for true explanations. Moreover, truth is not an all-or-nothing affair, and it is widely accepted that intentional distortions of truth are sometimes needed when constructing explanations.

We accept the basic realist view of the factivity of explanation because totally false explanations would provide incorrect answers to *what if* -questions and would thus provide only illusory understanding. However, the really interesting comparative dimension is introduced when two explanations incorporate or presuppose different numbers and degrees of falsehoods. One explanation is factually more accurate than another if it has (roughly) the same level of abstraction and detail, but includes fewer falsehoods. An especially important kind of falsehood is idealization. By idealization we mean the intentional distortion of the values of some relevant factors to some easily managed extremes (such as 0,1, or infinity) (Jones 2005). *Ceteris paribus*, a factually more accurate explanation enables a broader range of correct

inferences than an explanation incorporating idealizations, presuming that the inferences can actually be drawn without the idealizations in question.

Many explanations involve some idealizations in the *explanandum* as well as in the *explanans*. They are addressing stylized *explananda* that are not accurate representations of the concrete events. In these cases some characteristics of the event have been idealized in order to make the central explanatory factors more salient or tractable. For example, most illustrations of Newtonian mechanics usually idealize away the unmeasured but in reality non-negligible forces like friction or air resistance in the conceptualization of the *explanandum*.

If the idealizations of competing explanations are about the different elements of the explanation, comparing the explanations can be difficult. In such cases, what matters is not the overall level of idealization, but the judgments about the importance of the idealized factors. As in the case of sensitivity, the crucial factors are the epistemic aims and the pragmatic context of inquiry. In general, the explanatory power gained through the introduction of suitable idealizations can easily outweigh modest inferential gains linked to factual accuracy of a relatively unimportant variable. Idealizations are often necessary when trying to improve the degree of integration or the cognitive salience of an explanation.

It is usual to appeal to criteria such as simplicity or elegance when arguing for a hypothesis, especially in cases in which there is no straightforward empirical way of discriminating between rival explanations of a phenomenon. We argue next that to the extent such appeals actually have anything to do with explanatory virtues, they have to do with two different kinds of considerations: the degree of integration and cognitive salience.

4.4. The degree of integration

The degree of integration into existing knowledge is our fourth dimension of explanatory power. Many see unification or connectedness to a larger theoretical framework as a major virtue or even the most important explanatory virtue. Some (e.g. Michael Friedman, Philip Kitcher) have argued that unification in fact constitutes explanatory understanding. In our view, these suggestions are based on an attribution error. Even though some of the most important explanatory advances in the sciences have contributed to theoretical unification, it does not follow that they are *explanatory* for this reason. Also related is the confusion between evidential and explanatory virtues. Coherence and increased unification of a belief system are sometimes thought to increase the degree of confirmation of the belief system. Less controversially, if an explanation is consistent with or follows from a well-supported theory, then it is more credible than an unconnected *ad hoc* explanation conjured out of thin air. Whatever epistemological merits these ideas may have, they are in any case distinct and independent of questions of explanatory virtues proper.

However, there is a sense in which an explanation that is integrated into a larger body of knowledge contributes to explanatory understanding more than do these bodies of knowledge considered separately. An integrated body of knowledge is more than the sum of its parts. The underlying reason is this: when an explanation is well integrated into a larger theoretical framework, the theoretical connections can expand the range of answers to different *what if* -questions in two ways. First, because of the inferential connections to an already existing body of knowledge, dependencies between factors in the background theory and different aspects of the *explanandum* phenomenon may open up unforeseen dimensions in which contrastive *what if* -questions concerning the *explanandum* can be answered. Second, the explanation itself may bridge previous gaps within the existing theory and thus enable answers to new *what if* -questions not directly concerning the original *explanandum* phenomenon.

Note that this account of integration requires that integration be more than formal compatibility or conceptual coherence. The bodies of knowledge should come together in a manner that allows new relevant inferences to be made about the phenomenon. The requirement of relevance makes the idea of integration a local notion: it depends on one's epistemic aims. This contextual notion is quite different from the global notion of unification advocated by the supporters of unification accounts of explanation.

Higher integration is usually possible only when local contingencies affecting the *explanandum* are abstracted or idealized away. Thus, detail and factual accuracy often have to be sacrificed in order to gain a greater degree of theoretical integration.

In some instances (but by no means all), a higher degree of integration is achieved only by compromising cognitive salience. This trade-off usually results from the fact that fitting a large number of phenomena under a single conceptual scheme usually requires concepts and inferential connections that are novel and unfamiliar. This renders previous heuristics and thought patterns inapplicable. This trade-off is most striking in developments of fundamental physics. This is why local and somewhat *ad hoc* explanatory models are often used, even though in principle more general equivalent alternatives might be available.

4.5. Cognitive salience

How easily a given explanation may be grasped is often dismissed as an explanatory virtue proper on the basis that such matters are pragmatic and do not have any serious philosophical bearing. This is a mistake. First, since the power of an explanation is always dependent on the range of counterfactual inferences that the explanatory information enables, the kinds of inferences possible for limited cognitive systems such as humans directly affect what can be

explained and understood by such cognitive systems.[13] Cognitive salience refers to the ease with which the reasoning behind the explanation can be followed, how easily the implications of the explanation can be seen and how easy it is to evaluate the scope of the explanation and identify possible defeaters or caveats. Some of the factors underlying cognitive salience are species-relative, and some of them are highly personal, but many are related to disciplinary traditions and training.

Second, many criteria according to which explanatory power is attributed are based on cognitive salience, although this may not always be obvious at first glance. Unification achieved through the use of common patterns of inference (as in Kitcher 1993) or the use of the same words referring to similarly structured forms of causal interaction in different fields can sometimes make it easier to transfer understanding gained in one domain of phenomena to another. This aspect of cognitive salience also accounts for the (limited) appeal of views according to which rendering the *explanandum* familiar would be constitutive of understanding (e.g., Weber and van Bouwel 2007). Explanations formulated using familiar terminology or argument structures are often cognitively salient, and familiarity is therefore often *indirectly* related to understanding conceived as the ability to make counterfactual inferences. Of course, many times this kind of explanatory virtue of unification is illusory or comes at a high price in terms of other dimensions of explanatory power. For example, the use of economic (rational choice) models in intuitively non-economic domains of phenomena is often argued for on the basis of the virtue of unification. The opponents of economics-inspired political science do not claim that all or even most explanations offered by rational choice models are false, but that the explanations that are actually supported be data are lacking in other qualities that would make them interesting. Most rational choice explanations in political science resort only to

[13] In this paper we discuss only individuals as cognitive agents, but our principal points also apply to (socially) distributed cognition and to group cognition. However, in these cases the constraints of cognitive salience can be of different kinds. Cognitive salience is an issue in which empirical sciences of cognition could make an important contribution to the theory of explanation.

"thin rationality", i.e. that the agents are simply pursuing (consistently) whatever it is that is in their interests, and are consequently about very general and abstract forms of social institutions and interaction. According to the critics, this feature makes these explanations consistent with just about every empirical phenomenon imaginable and thus lacking in inferential power. On the other hand, the explanations that do require thick rationality, i.e. an account of what the agents' interests actually are, are often little more than formalizations of pre-theoretical folk-psychological guesses and are thus, according to the critics, banal. (Green and Shapiro 1994.)

Cognitive salience is not the same as simplicity. Maximizing formal simplicity by reducing the number of axioms, rules of inference, entities or kinds of entities might not be the best strategy for improving cognitive salience, although adding complexity often decreases cognitive salience. Cognitive salience is about the ergonomics of cognition, not about the formal properties of theories. Two formally equivalent versions of a theory might be very different from the point of view of their use in making relevant inferences about counterfactual situations. Von Neumann's Hilbert-space axiomatization of quantum mechanics may be formally elegant and economical, but in practice more cumbersome formulations of Schrödinger and Dirac more often used precisely because they are actually easier to use (Humphreys 1993). Although the explanatory facts are objective and independent of our reasoning abilities, the way in which they are represented is highly relevant from the point of view of understanding and communicating them.

5. Mechanistic detail and causal importance

The idea that unification has something to do with explanatory power was shown above to be explainable from our perspective, at least to the extent that it is genuinely linked to explanatory virtues. We will now discuss two further criteria – causal detail and causal importance - used to

evaluate explanations and show how these criteria too can be accounted for on the basis of our dimensions.

People prefer detailed explanations to mere sketches of explanation. This shows that detail is regarded as a virtue in explanatory contexts. An explanation is more detailed when it omits less of the relevant information. The emphasis here should be on the word "relevant". There has been some discussion in the philosophy of science about whether the addition of irrelevant details makes the explanation completely unexplanatory or whether it just makes an explanation worse (see Salmon 1998). We do not need to take a stance on this partly verbal issue. In this context it is sufficient to point out that irrelevant details at least decrease the cognitive salience of the explanation.[14]

It is clear that people often regard programmatic or sketchy explanations with suspicion. However, it is not obvious whether detail is valued as an evidential or an explanatory virtue. In our view it has both characteristics. Let us begin with evidential virtues. Many people consider detailed explanations more likely to be true. Of course, there is no direct link between detail and likeliness: fairy tales can be also extremely detailed. However, more detailed hypotheses are more easily shown to be false. If the details of the hypothesis have survived critical scrutiny, then one might regard that hypothesis as being more broadly supported by the evidence. Broad evidential support is obviously an epistemic merit, so detail is clearly related to evidential virtues. Of course, here again one must remember that we are talking about relevant details: irrelevant details do not add support to the central hypothesis.

What is the explanatory value of mechanistic causal detail? In our view, it is not a separate and independent source of explanatory virtue: its value can be explained in terms of other, more fundamental explanatory virtues. We take the addition of relevant details to mean the addition

[14] Possible exceptions to this claim are cases in which details that are explanatorily irrelevant nonetheless enhance the ability to recognize or recall explanatory information. However, even in such cases, it is important that the person is able to distinguish between these additional factors and proper explanatory factors.

of information concerning the mediating mechanism connecting the *explanandum* and the *explanans*. The view that the elaboration of the mediating mechanism improves an explanation is widespread, but the explanatory role, as opposed to the evidential role, of mechanistic detail is usually left unclear. In our view the value of explanatory mechanisms lies in the fact that underlying every *why*-question is a set of *how*-questions. We do not merely want to know why something happened; we also want to know how the cause brought about the effect. The details of the causal process, the mechanisms that account for why the causes had the effect they did, and even the presuppositions about the explanation can all contribute to understanding these questions. Understanding how a cause brought about its effect means simply the ability to answer more *what if* -questions concerning factors that were originally omitted from the explanation, such as what would happen if some parts of the mechanism were altered.

In principle, if we restrict our attention only to the original *explanandum*, then the explanatory import of mechanistic detail can be accounted for by the two virtues of increased precision of the *explanandum* and better knowledge of the sensitivity of the *explanans*. First, the added detail in the *explanans* can make it possible to account for a sharper *explanandum* or it can enhance the range of contrasts. Added detail can also provide more answers to possible aspects or contrasts in the original *explanandum* by creating totally new contrast-dimensions.

Second, the added details can enhance an explanation by adding information about the sensitivity of the explanatory link. When we know more about the mechanism transmitting the causal influence, then we have a better idea of what kinds of factors can disrupt the causal link and how. With the help of this information, we can answer more *what if* -questions concerning situations in which the background assumptions or conditions are different. Often the increase in understanding conferred by the introduction of mechanistic detail is due to an ability to answer *what if* -questions concerning factors other than the original *explanandum*. Thus, knowing more about how the cause brought about its effect often broadens the explanatory landscape rather than improving the original explanation.

However, more detailed explanations are not always better. First, the addition of extra details can compromise the cognitive salience of the explanation. A more detailed causal story might be more difficult to grasp. This is a simple fact of human cognition; our memory and ability to focus are limited and burdening them restricts our inferential performance. In such cases our ability to answer *what if* -questions will decrease. However, the interaction between detail and cognitive salience does not always work this way: the added detail can also enhance our inferential ability. The missing details might be crucial for the applicability of our mental models.

The second way in which the lack of detail might be a virtue is related to the role of abstraction in scientific thinking. By abstraction we refer to the intentional omission of factors known to be relevant (Jones 2005). Abstraction can increase but not decrease the scope of an explanation sketch: if details are removed from the explanation, then the explanation is applicable to a potentially larger set of phenomena. The increased level of abstraction can also enhance the level of integration and cognitive salience. These are clearly virtues that are relevant in theoretical contexts. Abstraction is also often necessary when we are ignorant of some relevant factors or when we cannot incorporate these factors into the explanation in a tractable manner, but we are not willing to sacrifice the virtue of the factual accuracy of the explanation.

Finally, we would like to make an observation about the relation between explanatory power and causal importance. Although our discussion has concentrated on causal explanations, we have left out a seemingly important dimension of explanatory power, namely, causal importance. It is commonplace to argue for the value of a preferred explanation on the basis that it picks out the most important cause. For many people, causal importance seems to differ from the other dimensions of explanatory power by being based on the notion of causal power. Causal powers are regarded as an objective property of the causal system of interest, and it is

thought that they can be characterized without reference to our theoretical or pragmatic interests.

However, the idea that there is such a thing as an objectively measurable (or at least comparable) causal power is, in most contexts, groundless. This idea is usually a result of an overly metaphysical and reified conception of causation as a ”thing”. In the majority of cases, different causal influences relevant to a given *explanandum* do not share a common currency in which their relative importance could be evaluated (Sober 1988). The Newtonian vector addition of forces is an extreme exception rather than the norm. Statistical exercises such as multiple regression and analysis of variance often create the impression that the regression coefficients or measures of fit such as the R^2 could be used as a common measure of causal strength of independent variables, but the appearance of additivity and separability of influences is usually an artefact of modeling assumptions rather than a property of the modeled causal system. Even in cases in which the causal assumptions required for a straightforward causal interpretation of regression coefficients or corresponding structural parameters as presented by Judea Pearl (2000) or Woodward hold, the causal effect of a variable is relative to the variation in all variables actually present in the population and may therefore be a poor guide to causal importance in any intuitive sense. For these reasons, in most cases the judgments of causal importance are based on judgments of explanatory power as defined in the dimensions presented above rather than the other way around.

6. Conclusions

In this paper we have sketched an account of explanatory virtues. We have argued that there are five separate dimensions and that these dimensions are sometimes in conflict. Apart from this descriptive aim, we have also tried to explain what makes these explanatory virtues epistemically desirable. This is our first stab at this unexplored, but important, topic in the

theory of explanation. Much remains to be done. It might turn out that there are some dimensions that we have missed or that we should make more fine-grained distinctions within the dimensions we have presented. It also remains to be shown whether this conceptual apparatus is actually useful in making sense of scientific controversies. Finally, although appeals to explanatory power are all too common in philosophical disputes (one's favorite position gives a better explanation of some conceptual puzzle or recalcitrant intuitions), it is not clear whether the concept of explanation used here is applicable to philosophical explanations, if indeed there are such things. We believe this problem highlights a grave deficiency in philosophers' methodological self-understanding. If appeals to explanatory power are used as philosophical arguments, then the implicit conception of explanation used and the associated standards of explanatory goodness should be made explicit. We have at least tried to provide a skeleton for any further developments in these problem areas.

The theory presented in this paper has some interesting implications for the debates about scientific explanation. First, the theory can explain what is intuitively appealing in mechanistic accounts of explanation that have become popular in the philosophy of science. Although it is a clear advance from the earlier covering-law account, the situation in the theory of explanation is still highly unsatisfactory. The accounts of mechanistic explanation do not provide any principled account of why mechanistic explanations are valuable, nor do they provide many suggestions for how mechanistic explanations are evaluated. Our account can help with both of these tasks. It can explain why explanations detailing causal mechanisms are epistemically valuable, without making mechanisms a necessary element of all explanations. And it can explain what kinds of mechanisms are valuable and by which criteria the goodness of mechanistic explanations is evaluated. This is a clear improvement over the current situation in which a mere metaphor (mechanism) serves as the core of an account of explanation. In addition, our account can explain why so many people mistake unification for explanation. As

discussed above, confusing evidential and explanatory virtues, different dimensions of explanatory power and merely intuitive notions of understanding can lead to an attribution error that misconstrues unification for explanation or regards unification as the central virtue of explanations. Finally, our account can be used to characterize differences in explanatory ideals of various scientific disciplines. Once these differences are made explicit, there is hope that controversies about these ideals could become more constructive, as our account offers some conceptual tools to articulate disciplinary ideals and commitments. The disciplinary explanatory ideals are no longer a god-given element of one's academic tribal identity, but variations in themes that all tribes share while having learned to emphasize them differently. Together these applications provide indirect support for our account of explanation: the contrastive-counterfactual account can do things that other accounts of explanation (the DN-account, unification account, causal process account, pragmatic account, etc.) have failed to do.

References

Baker, G. P. and Hacker, P. M. S. 2005: *Wittgenstein: Understanding and Meaning, Part I: Essays. 2nd. Rev. ed.*, Oxford: Blackwell.

Garfinkel, Alan 1981 *Forms of Explanation*. Yale University Press. New Haven.

Green, Donald P. and Shapiro, Ian 1994: *Pathologies of Rational Choice Theory: A Critique of Applications in Political Science*. New Haven: Yale University Press.

Hitchcock, Christopher and James Woodward 2003: Explanatory Generalizations, Part II: Plumbing Explanatory Depth, *Noûs* 37: 181-199.

Humphreys, Paul 1993: Greater Unification Equals Greater Understanding?, *Analysis* 53(3): 183–188.

Jones, Martin R. 2005: Idealization and Abstraction: A Framework, in Jones and Cartwright (eds.), *Idealization XII: Correcting the Model. Poznan Studies in the Philosophy of Sciences and the Humanities* vol. 86. Amsterdam/New York: Rodopi, 173-217.

Kitcher, Philip 1989: Explanatory Unification and the Causal Structure of the World, in Kitcher and Salmon (eds.), *Scientific Explanation. Minnesota Studies in the Philosophy of Science vol XIII*. Minneapolis: University of Minnesota Press, pp. 410-505.

Kuorikoski, Jaakko 2007: Explaining With Equilibria, in Persson and Ylikoski (eds.) 2007, 149-162.

Lipton, Peter 2004: *Inference to the Best Explanation. 2nd ed.* London: Routledge.

Miller, Richard W. 1987: *Fact and Method. Explanation, Confirmation and Reality in the Natural and the Social Sciences*. Princeton: Princeton University Press.

Morton, Adam 2002: *The Importance of Being Understood: Folk Psychology as Ethics*, London/New York: Routledge.

Pearl, Judea 2000: *Causality – Models, Reasoning and Inference*, Cambridge: Cambridge University Press.

Persson, Johannes and Petri Ylikoski (eds.) 2007: *Rethinking Explanation.* Dordrecht: Springer.

Salmon, Wesley 1998: *Causality and Explanation*. Oxford: Oxford University Press.

Schaffer, Jonathan 2005: Contrastive Causation, *The Philosophical Review* 114: 297-328.

Sober, Elliot 1988: Apportioning Causal Responsibility, *Journal of Philosophy* 85, 303-318.

Weber, Erik and van Bouwel, Jeroen 2007: Assessing the Explanatory Power of Causal Explanations, in Persson and Ylikoski (eds.) 2007, 109-118.

Wittgenstein, Ludwig 1953 [1997]: Philosophical Investigations (translated by G. E. M. Anscombe) Oxford : Blackwell.

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

Woodward, James 2006: Sensitive and Insensitive Causation, *The Philosophical Review* 115: 1-50.

Ylikoski, Petri 2005: The Third Dogma Revisited, *Foundations of Scienc*e 10, 395-419.

Ylikoski, Petri 2007: The Idea of Contrastive *Explanandum*, in Persson and Ylikoski (eds.) 2007, 27-42.

Ylikoski, Petri 2009: Illusions in Scientific Understanding, in De Regt, Leonelli & Eigner (eds.) *Scientific Understanding: Philosophical Perspectives*, Pittsburgh University Press, forthcoming.