Published 2009 in *Scientific Understanding: Philosophical Perspectives* (edited by H. De Regt, S. Leonelli & K. Eigner), Pittsburgh University Press: 100-119.

The Illusion of Depth of Understanding in Science

Petri Ylikoski
Department of Social Research
University of Helsinki
petri.ylikoski@helsinki.fi

1. Introduction

Philosophers of science have a long tradition of making a connection between explanation and understanding, but only lately have they started to give the latter notion a substantial role in their theories. The reason is easy to understand: understanding is an even more difficult notion than explanation. To my mind, the recent interest in understanding (exemplified by this volume), springs from something important: explanation is a cognitive activity, and for all too long, theories of explanation have dismissed the cognitive dimension on the lame excuse of its being a too 'subjective' ingredient for a theory of *scientific* explanation. Explanation is connected with understanding and we just have to deal with it. But how should we do that?

In this chapter I will employ a well-known scientific research heuristic of studying how something works by focusing on circumstances in which it does not work. Rather than trying to describe what scientific understanding would ideally look like, I will try to learn something about it by looking at mundane cases where understanding is partly illusory. The main thesis of this paper will be the following: scientists are prone to the illusion of depth of understanding (IDU), and as a consequence, they sometimes overestimate the detail, coherence, and depth of their understanding. I will start my argument by presenting an analysis of the notion of understanding and its relation to a sense of understanding. In order to make plausible the claim that these are often disconnected, I will describe an interesting series of psychological experiments by Frank Keil and co-authors. These experiments suggest that ordinary people routinely overestimate the depth of their understanding. In Section 3, I will argue that we should take seriously the possibility that scientific cognition is also affected by IDU. Section 4 will spell out some possible causes of explanatory illusions in science. In the final section, I will discuss how scientific explanatory practices could be improved and how the philosophy of science might be able to contribute to this process.

2. Understanding and the Illusion of Depth

Let us start with the notion of understanding. What is it? I agree with Wittgenstein when he argues that understanding should not be understood as a sensation, an experience, or a state of mind. It is not primarily a process: coming to understand something is a process, but not understanding itself. More generally, understanding is not a special moment or phase, but a more permanent attribute. It is an ability. When a person understands something, she is able to do certain things. (Wittgenstein 1953, §§143-155, 179-184, 321-324; Baker and Hacker 2005, 357-385.)

This does not mean that understanding is some sort of special skill. Understanding consists of knowledge about relations of dependence. When one understands something, one can make all kinds of correct inferences about it. Many of these inferences are counterfactual: What would have happened if certain things had been different? What will happen if things were changed in a certain manner? To get a better grasp of these counterfactual inferences, it is useful to consider the ways in which one can demonstrate that one understands something.

Let us start with something practical. In favorable circumstances understanding allows for successful *interaction* with an object. One's knowledge of the relevant relations of dependence allows one to make inferences about the consequences of one's interventions. One knows what can be done and how. One also knows what cannot be done. In this way, understanding gives an ability to *control* the phenomenon. Understanding comes in degrees and the amount of control can be used as its measure: other things being equal, the wider the range of control, the deeper is one's understanding. With enough understanding of how something works, one can *repair* it when it fails or even *build* a new one. Again, the more failures one can repair, the better is one's understanding of it. A similar point applies to understanding human beings. The better one understands a person, the more successful one is in cooperating and communicating with her.

One is not always in a position to causally interact with the phenomenon, so the amount of control cannot be regarded as the ultimate criterion for understanding. If the fact to be understood is in the past, or if we lack means of intervention, we cannot demonstrate our understanding in this way. But there are alternatives.

If causal interaction with an object is not possible, we might still be able to demonstrate our understanding by *anticipating* how the object behaves. Here the relevant inferences would be about the future consequences of some event or series of events. In this case we would not make inferences about the consequences of our imagined or real interventions, but would predict what will happen. Some of these predictions are about the real world, and others are

about counterfactual situations. Again, the amount of understanding can be roughly measured by the scope of one's predictive abilities. Other things being equal, the broader the range of circumstances one's anticipatory understanding covers and the more precise one's predictions, the better is one's understanding of the phenomenon. However, other things are not always equal, so one cannot equate the amount of explanatory understanding with the ability to predict. Anticipatory understanding is just one criterion for understanding.

One might be able to interact with an object or predict how it will behave without being able to make this knowledge explicit, which is required in science. For this reason, the above criteria are not sufficient in the case of scientific understanding. Something more is required – the ability to explain, that is, to communicate one's understanding. This constitutes the third way to demonstrate understanding. Tacit understanding might give the ability to make predictions and to interact with an object, but in science it is required that one be able to make one's knowledge explicit by stating the underlying principles. Let us call this *theoretical understanding*.[i] Scientific understanding is demonstrated by giving explanations, that is, by communicating one's understanding. Explanations are answers to questions of why, more specifically, they are answers to questions of *what if things had been different* (Woodward 2003), so again we are dealing with an ability to make counterfactual inferences. We also have a rough measure of the amount of understanding: the more, and more diverse, explanation-seeking questions one is able to answer, the better (or deeper) is one's understanding. In this case explanatory knowledge and understanding go hand in hand.

These three ways to demonstrate one's understanding constitute criteria according to which people attribute understanding to each other.[ii] They are concerned with external displays of the knowledge one has. This knowledge might be tacit or explicit, the crucial thing is that it gives an ability to do certain things. Some authors suggest that understanding involves having an internal mental model of the object of understanding (for example, Waskan 2006). This is an interesting suggestion, but it cannot be the whole story about understanding. When we evaluate someone's understanding, we are not making guesses about her internal representations, but about her ability to perform according to standards we have set. The concept of understanding allows that the ability can be grounded in various alternative ways, as long as the performance is correct. Furthermore, the correctness of the internal model is judged by the external displays of understanding, not the other way around. This makes understanding a behavioral concept.

It is plausible that having an internal mental model is a causal condition for understanding, but it should not be equated with understanding. It would also be a mistake to commit oneself too tightly to the idea of internal representation – much of scientific cognition employs various

sorts of external representations. Successful scientific cognition combines internal representations with external representations and other tools. This fact is easily forgotten when one concentrates on mental models. The distinctive feature of scientific cognition might not be some special mental models but the extensive use of various external representations, as recent work on distributed cognition suggests (Donald 1991; Hutchins 1995; Clark 1997; Giere 2002).

If understanding is an ability rather than a mental state or an experience, why do many people think otherwise? The reason is that there is a mental experience that is closely related to understanding: *the sense of understanding*. It is a feeling that tells us when we have understood or grasped something. Now it is an open question whether this sensation is always the same, but in any case we recognize it when we have it. This sense of confidence and the feeling that (often) comes with it can be easily confused with what we think it indicates: understanding. Ideally these two things would go hand in hand, and assimilating them should not create any trouble. Real life is different. The sense of understanding is only a fallible indicator of understanding (Grimm this volume). Sometimes one has a false sense of understanding and sometimes one understands without having any associated feelings or experiences.[iii]

The existence of the sense of understanding should not be regarded as any kind of oddity. It is a special case of feeling of knowing, a much discussed *metacognitive* notion. According to Korian (2000) the feeling of knowing serves as an important source of information in our cognitive activities. For example, it helps to regulate our learning and recall activities. Frank Keil (2003) suggests that the sense of understanding has a similar metacognitive role. It gives us confidence to try things, and when it is lacking we can sensibly abstain from the activity in question. It also guides the search for new knowledge. It tells us when to stop the search for new information; it signals when we know enough. A stopping device like this is very useful. A comprehensive understanding of something would require huge amounts of time and effort, and might still not be achievable. In everyday life, this kind of ideal knowledge is impractical. It is better to have some signal that tells us when we have enough knowledge to function effectively, even if this indicator is not wholly reliable. In addition to this, the sensation associated with the sense of understanding can have a motivational role. Satisfying curiosity is highly rewarding (Schwitzgebel 1999, Gopnik 2000, Lipton this volume). It provides motivation for learning and other cognitive activities. In this way it provides a psychological mechanism that has an important role in human cognition. The desire to satisfy one's curiosity also provides an important psychological motivation for doing scientific research.

Although the phenomenology of the sense of understanding can mislead one into thinking that understanding is an on-off phenomenon ('Now I got it!'), it actually comes in degrees. First, the understanding can be about different aspects of the phenomenon. Second, those aspects may be understood in various degrees. Consider an ordinary object like a personal computer. Different individuals understand to varying degrees how their computer works. Some might know about the software, or some specific piece of software, and others the hardware. Most people just use the software without any understanding of the internal workings of a computer. Despite these differences, they all understand something about their PC. The crucial question is what aspects of it they understand. By asking *what* has been understood, the extent of understanding *can always be specified*. There is nothing mysterious in the idea that understanding comes in degrees, nor in the idea that there are clear differences in the degree of understanding that different individuals possess.

From the point of view of this paper, the crucial thing about the sense of understanding is that it is a highly fallible source of metacognitive information. It does not give us direct access to knowledge that is the basis of our understanding. Like any other metacognitive judgment, it can misfire: a false sense of understanding is a real possibility. In such cases one overestimates one's understanding. One can also underestimate one's understanding. In these cases one thinks that one understands less than one actually does. In fact, it would be highly surprising if the sense of understanding would turn out to be perfectly calibrated with our understanding.

The fallibility of the sense of understanding can be demonstrated experimentally. People often overestimate the detail, coherence, and depth of their understanding. Frank Keil calls this effect the illusion of depth of understanding (IDU). Together with his associates (Rozenblit and Keil 2002; Mills and Keil 2004), he designed an experimental set-up where participants were first taught how to use a seven-point scale that rates their knowledge. They were then asked to rate how well they know how various devices or systems work. After rating their understanding of a large set of these items, the participants were then asked to explain in detail the actual workings of these systems. After giving each explanation, they were asked to re-rate the depth of their initial knowledge. The participants were then asked to answer diagnostic questions that experts consider to be central to understanding the system, after which they were asked to re-rate their knowledge again. Finally, they were presented with a concise but thorough expert explanation and were asked to rate their initial knowledge once more in light of that expert explanation.

The results across several studies show a strong drop in ratings of knowledge after each re-rating and often the participants being shocked and surprised by their own ignorance. The

general conclusion is that most people are prone to feel that they understand the world with far greater detail, coherence, and depth than they really do. According to Keil, this effect is distinct from the general overconfidence effect found in many psychological studies. Within the experimental set-up described above, the illusion of having detailed and coherent knowledge occurs primarily for explanatory understanding. In contrast, people's ratings of how well they know facts, procedures, or narratives are quite well calibrated and they are not surprised at what they actually know. (Rozenblit and Keil 2002; Keil 2003.)

What is behind this interesting phenomenon? Keil suggests four possible contributing factors. One factor is confusion between what is represented in the mind with what can be recovered from a display in real time. People may underestimate how much of their understanding lies in relations that are apparent in the object as opposed to being mentally represented. A second factor may be a confusion of higher and lower levels of analysis. For example, while explaining how a car works, one might describe the function of a unit, such as the brakes, in general terms, and then describe the functions of subcomponents, such as brake lines and brake pads, which in turn can be broken down even further. The iterative structure of explanations of this sort may lead to an illusion of understanding when a person gains insight into a high-level function and, with that rush of insight, falsely assumes an understanding of further levels down in the hierarchy of causal mechanisms. A third possible factor is the fact that many explanations have indeterminate end states. One usually has little idea of what the final explanation will look like and the end state is largely indeterminate from the posing of the question. This makes self-testing one's knowledge difficult. The final factor in Keil's list is the rarity of production: we rarely give explanations and therefore have little information on past successes and failures. (Rozenblit and Keil 2002, 522-523.)

Keil's thesis about IDU looks quite similar to the claims made by J. D. Trout (2002), so it is important to see their differences. Trout argues that the sense of understanding is often influenced by the overconfidence bias and that this makes it a highly unreliable source of information. However, he does not cite any studies about explanatory cognition to support his argument. His argument against the epistemic relevance of the sense of understanding is based on the idea that humans are generally biased towards overconfidence. Lately this idea of general overconfidence bias has been criticized on theoretical and methodological grounds (Juslin, Winman, and Olsson 2000). If these criticisms are right, Trout's argument is in trouble. However, these criticisms do not apply to Keil's experiments. He is making a claim that there is a specific overconfidence effect in the assessment of understanding. The general overconfidence effect does not have any role in his argumentation. Similarly, his experimental

set-ups are not based on the assumptions that the critics of overconfidence research have found problematic. Finally, the conclusions he draws are different: he does not suggest that we should give up the notion of understanding, as Trout does.

Keil's studies show that understanding and the sense of understanding do not always go hand in hand. Could the sense of understanding be calibrated to be a more reliable indicator of understanding? We simply do not know. There are no empirical studies of possible ways of improving explanatory practices. It is possible that if we focus more on our explanatory practices, and make our explanations more explicit, the calibration of our sense of understanding would improve. This would help to address two causes of IDU suggested by Keil: the indeterminacy of the end state and the rarity of production. However, there are some grounds for being skeptical of our prospects in calibration. The extensive experimental literature on reading comprehension shows that the calibration of comprehension is not easily achieved (Lin and Zabrucky 1998).

3. The Case for Explanatory Illusions in Science

In what follows I will extend the claims of Keil and his associates to the domain of scientific research. Their studies concentrated on ordinary people's illusion of understanding, and they avoided discussing this in the context of scientific enquiry. The central claim of this chapter is that IDU is possible and indeed common in the sciences. Here I will give a number of reasons for this suspicion. My claims will be empirical hypotheses about factors affecting scientists. I will not claim that these problems are unavoidable, but claim that they are prevalent enough to be taken as a ground for taking seriously the possibility of IDU in science.

The first reason to suspect that IDU might be relevant to science is the continuity between scientific and everyday cognition. Scientists use the same cognitive mechanisms as everybody else. Although scientific standards for evidence, instruments and social practices make scientific cognition different from lay cognition, we should not assume that their use automatically makes the problems related to evaluating explanatory understanding disappear. Most of the things that contribute to the differences between scientists and ordinary people are not related to the assessment of explanations.

The second reason is the level of attention the articulation and evaluation of explanations gets in the scientific community. Although many scientists and philosophers are enthusiastic about explanation as a proper cognitive aim of scientific enterprise, it is surprising how little explicit concern explanatory practices get in the sciences. As an example, consider scientific publications. Scientific journals provide many guidelines for presenting the data and the

methods. However, they do not provide any direction for presenting explanations. Typically, scientific journals are structured to report empirical results, and as a consequence, they do not put much emphasis on explanatory practices. The explanatory claims are located in the discussion section, and they are often left implicit or vague, and the editors are often happy with this. Of course, not all scientific publications follow this format nor do they all concentrate on reporting experiments or observations, but in general they still share the same lack of explicit focus on explanation. There are no specialized scientific forums to discuss or evaluate explanatory claims. Nor is the situation better in science education. What students learn about presentation and evaluation of explanations, they learn implicitly. Scientific education does not contain formal instruction on explanatory practices. Although optional philosophy of science courses might contain some discussion about explanation, the content of these courses might still be more or less irrelevant from the point of view of the explanatory challenges the students will face.

My third reason to expect IDU to be relevant is based on the fact that the circumstances of scientists are more difficult than those of the participants in Keil's experiments. Scientists cannot use correct expert opinion as a benchmark, and the criteria for the correctness of the explanations are themselves ambiguous. This means that the end state is even more indeterminate, making self-testing more difficult.

The fourth reason is the role of the division of cognitive labor in science. It is commonly recognized that science is a collective enterprise. The subject of scientific knowledge is not an individual scientist but a community characterized by an extensive division of cognitive labor. In this context it is not possible to spell out the various ways to understand this idea, but fortunately this is not needed to make my point. No matter whether the scientist is trying to assess what she herself, her research group, or her discipline, or science as a whole really understands, she is involved in evaluating understanding that is based on a division of labor. This means that she must evaluate not only her own knowledge, but also other people whose competence is different from her own. This creates additional difficulties for assessing understanding. First, she cannot rely on her sense of understanding, as it does not give any kind of access to other people. Second, she is forced to evaluate the people (and theories) outside her narrow special field based on relatively shallow understanding; basically she is not competent to evaluate them. For this reason, she must employ various indirect indicators of competence, like reputation and the general standing of the field. Furthermore, she is not often in a position to test the understanding of the other persons, so she is more or less forced to simply accept their word for their competence. The situation is further complicated by the fact

that there are various incentives for scientists to overstate their understanding to others. Everybody wants to impress, and intellectual authority is a valuable resource. For all these reasons, a scientist might be too impressed about the understanding of others and consequently might be prone to overestimate the collective level of understanding of the scientific community.

The fifth reason is the future-oriented nature of scientific evaluation. Most of the time scientists evaluate their theories, models and methods in the context of making decisions about their future research. Although most of the philosophical discussion about the problem of theory choice is set in the context of acceptance, scientists make their choices mostly in the context of pursuit. They are not primarily assessing the current achievements of their theories, but making informed guesses about their fruitfulness in the future. In other words, they are not choosing which theory to accept, but choosing which theory they are going to work with. (The same point applies to scientists serving as evaluators for funding agencies: they try to pick the most promising plans, which are not necessarily the ones that are currently best supported by the evidence.) In the context of pursuit they do not have to make up their minds about acceptance, but they must place their bets on the most fruitful approach. This observation is also highly relevant to our discussion. Scientists assessing explanations are not simply evaluating their current status in terms of explanatory virtues and evidential support; they are also making guesses about the future prospects of these explanatory hypotheses. The real issue is the promise of future understanding, not the things that have been delivered thus far. This future orientation makes the product to be evaluated extremely ambiguous. The point in time when the evaluation is supposed to take place is not specified. Furthermore, the future versions of hypotheses to be evaluated cannot be spelled out, so one does not really know what one is evaluating, nor does one know what one is evaluating it against.[iv] If assessing existing explanations is difficult, the assessment of future explanatory potential might just be impossible.[v] Of course, the evaluation is not completely in the air. The compatibility with other relevant theories is one criterion; another is the past track record of the candidate. But these are still indirect criteria, and people can reasonably disagree on how to use them.

I submit that, most of the time, scientists make their choices on much more pragmatic criteria than the promise of explanatory excellence. They choose opportunistically what is workable: the line of work that promises doable problems that can be solved by the expertise, skills and research resources they have at their disposal. These primary reasons do not have much to do with explanation, but the scientists might still attempt to justify their choices (for example, to possible funding sources) by appealing to future epistemic virtues. This would make

explanatory argumentation more or less mere window-dressing. The real reasons for the choices might be different from the ones that are publicly presented.

4. Seven Sources of IDU

The previous section already indicated reasons to expect that IDU is a real danger in science. In this section I will investigate some factors that might make scientists prone to *miscalibrated* assessments of the depth of their understanding. My suggestions will be a bit speculative, but this is unavoidable, as I have not found much empirical research that would be relevant to this issue. However, the individual hypotheses that I will present are all empirically testable. I do not claim that the following list is exhaustive; there are probably also other relevant sources of IDU. It is also possible that their relevance may vary in different scientific fields. I will not make guesses about how strong an influence these factors might have, nor I am making judgments about their relative importance. As well, I leave the direction of the miscalibration open: given the point made in the previous section, it is probable that scientists are disposed to *overestimating* their understanding, but it is also possible that some of the following cause an *underestimation* of the understanding (in some circumstances). All I wish to do is to strengthen the case for taking the possibility of IDU in science seriously.

The first possible source of IDU is the simple *failure to find things puzzling*. It is typical in everyday cognition that only anomalous or otherwise surprising observations raise the need for an explanation. Only the unexpected or abnormal events challenge one's existing scheme of categorization. The same is probably also true for the cognitive life of scientists, at least most of the time. Of course, it is an important characteristic of science that it also asks why questions about the normal (Ylikoski 2007), but the investigative focus is always directed at a limited number of issues, so this difference does not change the basic point I wish to make. How does a failure to find things puzzling contribute to IDU?

A puzzling phenomenon forces one to come up with an explanation and at the same time to face the limits of one's understanding. This situation is in sharp contrast to familiar things that behave as they are expected to. They do not call for explanation, and as a consequence, do not challenge a person's confidence in the depth of her understanding. The ability to find a categorization for the initially surprising observation brings about a sense of confidence in one's conceptual scheme. This restoration of confidence in the conceptual scheme might be confused with an increase in understanding and regarded as a consequence of explanatory insight. In this way, familiarity can sustain an instance of IDU. But it can also give rise to one. The puzzling things that are not understood are contrasted with the familiar and the expected.

This contrast can easily be mistaken for a contrast between being understood and not being understood. Furthermore, an explanation aims to remove the puzzlement about the phenomenon, so understanding an anomaly makes it appear normal. From this it is easy to infer that the observations regarded as normal already have the property of being understood. Of course, this inference is not valid, but that does not mean that we can easily escape making it.

The second possible source of IDU is *the confusion between explanation and description*. Quite often one gets the impression that people think that they automatically contribute to the explanatory understanding of an event simply by finding out and describing facts about its causal history.[vi] Furthermore, they might think that the more facts about the causal history you have, the better is your understanding of the *explanandum*. This is an interesting confusion. Causal history is naturally relevant for explanation, but the crucial point is that not all facts about the causal history are relevant to a given *explanandum*. Their relevance depends on the aspects of the *explanandum* one is trying to make sense of.[vii] It might be that every single fact about the causal history is relevant to some explanation-seeking question about the *explanandum* event. But the fact is that we are not interested in all possible explanation-seeking questions related to the *explanandum*. We are not looking for questions that would make our findings explanatory, but for facts that would answer our explanation-seeking questions. Despite this, some scientists seem to be involved in this kind of 'explanatory fishing': they think that their findings are important because they might be crucial in explaining *something.* What constitutes that something is conveniently left unarticulated. This ambiguity helps to maintain the illusion of gained understanding. The illusion itself might be a product of the fact that the research is hard work, even if the findings turn out to be irrelevant: one likes to feel that one has achieved something, in this case, explanatory understanding.

The third possible source of miscalibration of the sense of understanding is *the ambiguity of the notion of explanation*. The first piece of evidence comes from the fact that despite extensive debate, philosophers of science have not reached a consensus about the nature of explanation. This is weak evidence, as typically philosophers cannot reach a consensus about anything. But in this case the philosophical disagreement reflects a more general uncertainty. Discussions with scientists show that they have quite different ideas about the criteria or characteristics of explanatory understanding. Sometimes they present an account that sounds quite strange to an interlocutor who is familiar with various philosophical theories, but sometimes scientists present ideas and concepts that they (or their teachers) have learned during a philosophy of science class. It seems that the notion of explanation is a kind of

metascientific notion that is not explicitly present in everyday scientific practice: the scientists manage to do their work without having an articulated or even shared notion of explanation.

When comparing explanations, scientists and philosophers often appeal to metaphors of explanatory power and depth. Very few have tried to articulate what these notions actually mean. However, there are grounds for thinking that these notions do not refer to a single attribute of explanations. For example, the notion of explanatory power can be used to refer to a number of different properties explanations. An explanation could be called powerful when it i) is cognitively salient (e.g. is easy to understand and use); ii) is factually accurate; iii) gives a precise characterization of the *explanandum*; or iv) is robust with respect to changes in background conditions (Ylikoski and Kuorikoski 2007). These virtues of explanation can sometimes be in conflict. This means that the evaluation of explanatory power is not one-dimensional. The important point in this context is that these different dimensions are not usually articulated in scientific practice and that it is probable that scientists confuse them, at least sometimes. Consequently, this confusion can lead to a miscalibration of the assessment of understanding.

The fourth source of IDU is another kind of ambiguity. One cannot explain anything completely; rather one always explains some aspects of the *explanandum* phenomenon. But which aspects are addressed with a given explanation? Quite often it is very hard to see what the precise *explanandum* is that scientists are trying to address with their theories. *The ambiguity of the explanandum* is a real problem that surfaces in scientific controversies. A similar point applies to *the ambiguity of the explanans*. All explanations take some things for granted and treat them as background assumptions. But which are dispensable parts of the background and which are the essential ingredients of the hypothesis? This is a difficult problem, even for the person suggesting the hypothesis. Scientists usually have intuitive ideas about the *explanandum* and the *explanans*, but they cannot fully articulate them. In these circumstances it is extremely difficult to say what has actually been achieved with the explanatory hypothesis and how much it was the hypothesis (in contrast to the background assumptions) that did the explanatory work. This gives more room for our wishful thinking and egocentric bias to operate.

The fifth source of IDU is c*ircular reasoning*. The identification of cases of circular reasoning is difficult outside formal logic (Rips 2002), and it is plausible that sometimes people can mistake a restatement or a presupposition of the *explanandum* for the *explanans*. In highly theoretical (and philosophical) contexts this fallacy might be quite common, as the concepts are often difficult to define precisely. Circular inferences are logically valid, and identifying

them is quite hard without making the whole argument explicit. Furthermore, as definitions of the concepts and structure of the argument are often ambiguous, judgments concerning circular reasoning are often controversial. This makes their identification and avoidance even more difficult.

The sixth possible source of IDU is *the confusion between the explanatory power of the hypothesis and evidential support for it*. When people are evaluating an explanation they are not only making judgments about its explanatory virtues, they also assess the reasons to believe it. Psychologists are quite confident that there exists an "explanation effect" according to which people estimate the probability of an event much higher when they have an explanatory story about it. A task, like giving an explanation, which requires that a person treats a hypothesis as if it were true, strengthens the confidence with which that hypothesis is held. People also seem to have difficulties in distinguishing between explanation and evidence. (Koehler 1991; Brem and Rips 2000.) Of course, this is a philosophical issue, as will be clear to anyone who has followed the discussions around Peter Lipton's (2004) distinction between likeliness and loveliness. Loveliness *might* be an indicator of likeliness, as many friends of inference to the best explanation believe. I do not wish to take a stand on this issue here. My point is that the mere existence of this debate is evidence for the claim that explanatory power and evidential support are difficult to keep separate. It is not always clear to which group a given criterion of explanatory assessment belongs. Notice that my suggestion here is a bit different from the claim made by psychologists. They are claiming that explanatory considerations influence the assessment of evidential support, whereas I am talking about influence in the other direction. If people do not explicitly distinguish between their confidence in the understanding provided by an explanation and their confidence in the truth of the hypothesis, it is possible that these will influence each other. As a consequence, a person might think that a hypothesis provides understanding because he is so confident that it is true.

The last source of IDU arises from the fact that scientists often use complex representations to make sense of the phenomena of interest. In order to understand the phenomenon, one must also understand the epistemic tools that are used as the medium of representation. In cases like this, understanding is mediated: an object is understood via a tool that also needs to be understood. How might this contribute to IDU? Learning to use the relevant mathematics, modeling assumption, and other ingredients of the representative medium requires a lot of work: acquiring the relevant abilities is not a trivial operation, but a real achievement for an aspiring scientist. At the same time, many of the objects that scientists are trying to understand are difficult to access. They are quite often abstracted, idealized, or stylized, in a manner that

makes it very hard to grasp them without the representative means provided by the theory. (Think, for example, of the phenomena most economic theories are trying to deal with.) In these circumstances it can happen that a scientist *confuses her understanding of the theory with her understanding of the phenomena* that the theory is intended to deal with. In other words, she confuses the intelligibility of the theory (De Regt and Dieks 2005) with the intelligibility of the phenomenon. This confusion is understandable, because in these circumstances the intelligibility of the theory is a necessary condition for understanding the phenomenon. Both involve understanding. However, understanding a theory is different from understanding a phenomenon, considering the ways in which the understanding can be demonstrated. The understanding of a theory is displayed by making the right sorts of inferences from it, by knowing what kinds of problems it can deal with and by building models, whereas the understanding of a phenomenon is ultimately displayed by making right predictions, successful interventions and by answering explanation-seeking questions about it. The latter cannot be done without a theory, but requires more than understanding the theory. After all, one can also understand theories that are considered false, outdated or irrelevant. Making sense of a theory does not guarantee that one makes sense of the phenomenon.

5. How to Avoid Illusory Understanding?

The above list of possible sources of illusions of explanatory understanding raises the question how seriously we should take explanatory ambitions of science. Is it possible to calibrate scientists' sense of understanding to a level that would make us confident enough in it? If it were possible, how could we improve our sense of understanding? Or if the answer is negative, is there any way around the bottleneck? I will not make any guesses about the answer to the first question. I hope future research may provide this. We should, however, not be too optimistic: the research on metacognitive self-assessment (Lin and Zabrucky 1998; Davis et. al. 2006) shows that the task is difficult and the case of explanatory understanding is probably more challenging than the cases of self-assessment these studies usually evaluate. I do not have ready answers to the two other questions, but I would like to suggest that the ways in which we can try to improve our sense of understanding also help us to gain some independence from it.

The key idea is to take seriously the idea that science is an intersubjective enterprise. The way science has succeeded in improving itself beyond our everyday individual cognitive performance has largely been due to three things: 1) the attempt to make the claims more explicit; 2) the use of external representations in this process; and 3) the development of social practices that allow critical debate on these representations. The quality of epistemic practices

has improved as we have given up the idea that everything must be done solely in one's own head (Donald 1991; Hutchins 1995; Clark 1997; Giere 2002). Maybe we could do something similar in the case of explanatory understanding. There is no reason to suppose that current scientific explanatory practices are as perfect as they can get. Most of the possible sources of IDU presented above are related to the intuitive manner in which judgments about explanatory adequacy are made. If the explanatory claims were made more explicit, their correct assessment would also be easier. This would provide better feedback for the calibration of the sense of understanding; moreover it would also make us less dependent on this quite fallible source of metacognitive insight. The crucial question is: How do we make our explanatory practice more explicit?

Here lies an opportunity for philosophers of explanation. They could adopt the aim of providing constructive suggestions for the improvement of scientific practices. In this way philosophers could prove their relevance to scientific enterprise and could acquire an external standard for evaluating the adequacy of their proposals. Currently, philosophers of science judge the competing theories of explanation mostly against their 'intuitions'. This is a standard procedure in the analytical philosophy, but it is not a very useful standard of evaluation. Philosophers tend to have very different 'intuitions' about explanation. Although individual's intuitions are strong and not easily changed, they seem to be products of an array of slightly suspect causes: the philosophy of science literature one has been exposed to during some sensitive period, the intuitions of people (philosophers or scientists) one takes to be authoritative, explanatory practices in the fields of science one is familiar with, one's personality, and so on. It would be better to have some external standard that is better connected with science practice.

It would not be an improvement to simply replace the philosopher's intuitions with intuitions of practicing scientists. Instead of fitting with intuitions, the philosophical theories should be judged on the basis of their adequacy in solving explanation-related controversies in the sciences and their usefulness in science education and communication, in short, in fighting IDU. Here the final judgment would rest with the scientists: they would evaluate which philosophical suggestions have proved fruitful in improving scientific practice. Philosophical theories should not simply describe current scientific practices, but suggest improvements. This methodological recasting could help to reformulate some old issues in the theory of explanation. I will give two brief examples.

The first example is the view Wesley Salmon (1998) has dubbed 'deductive chauvinism' and according to which explanations are deductive arguments. Although the argument that this

view is somehow tied to the assumption that the world is deterministic is a failure, the philosophical debate over the last decades has established that deduction does not have a constitutive role in explanation (Ylikoski 2005). People have been giving explanations long before the invention of formal logic, there are plenty of non-propositional explanations (for example, using diagrams and pictorial representations), and most propositional explanations do not have a deductive structure (although they might have some deductive passages). In this standard sense deductive chauvinism is dead, but I think we should not give up this idea so easily.

We should not regard deductivism as a claim about what makes explanations explanatory, but as a practical suggestion for improving explanatory practice. Explanations are not deductive arguments, but *they can be reconstructed as such* (Ylikoski 2005). The idea is that an (even partial) attempt at deductive reconstruction leads to improvements in the *process* of articulating explanations by forcing one to explicate both many of the background assumptions and the intended *explanandum*. I think that this is the main reason why some people are still intuitively attracted to the deductive ideal. Of course, deductivism is not a foolproof procedure for fighting IDU. If the source of illusion is circular reasoning, the deductive test is not helpful. (Naturally, deductive reconstruction might still make it more visible.) A more serious problem is formed by the various ways to fudge the explanation by using filler terms and other placeholders. Deductive reconstruction does not help with these.

The other example is the idea that the *explanandum* is contrastive (Woodward 2003; Lipton 2004; Ylikoski 2007). This old idea has usually 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 best ways to use the contrastive idea. The contrastive thesis can be regarded as a thesis about fruitful way of explicating the intended *explanandum*. All explananda can be reconstructed in contrastive terms and the suggestion is that this is helpful both from the point of view of enquiry and from the point of view of evaluating explanations. Understood in this way, the contrastive proposal includes two theses: one about reconstructability and another about fruitfulness. Clearly the latter is more interesting, as it might hold even if the first one does not hold universally. The idea is that the contrastive reconstruction helps to be more specific about the intended *explanandum* and in this way it makes the comparison of apparently competing explanations possible: quite often it turns out that these explanations address a slightly different aspect of the *explanandum* phenomenon.

These two suggestions are only examples. There are probably many other ways in which scientific explanations could be made more explicit in a manner that prevents us from becoming victims of IDU. Finding out these practical measures should be one of the chief aims of the theory of explanation. In this way it could contribute to the truly philosophical enterprise of figuring out the limits of our (current) scientific understanding. After all, explicitly spelling out our understanding often shows that we understand less than we originally thought.

6. Conclusion

In this chapter I have argued for a number of different theses about explanatory understanding. I first argued that understanding should be characterized as an ability and that it should be distinguished from the sense of understanding. This is a distinction that is not commonly made in philosophy of science literature. When understanding is analyzed as an ability to make counterfactual inferences in contexts of manipulation, prediction and explanation, its relation to knowledge can be clarified. Understanding is only knowledge about the relations of dependence, not something mysterious added to knowledge. This characterization allows for the possibility of tacit understanding, but the main focus of this paper has been on scientific understanding, that is, explanatory understanding. In my view, one has explanatory understanding when one is able to answer explanation-seeking questions. The more questions one can answer about a phenomenon, the better one's understanding. This makes the connection between scientific understanding and explanation quite close. However, this does not mean that these notions are equivalent. Explanations are best analyzed as answers to explanation-seeking questions. The notion of understanding is more appropriate for characterizing the epistemic aims of science: organized knowledge that allows one to answer to a whole series of what-if-things-had-been-different -questions about the world. In short, understanding is an ability to give explanations.

The second main theme in the chapter was the possibility of illusory understanding in science. The key here is the sense of understanding that has an important metacognitive role in our cognition. Despite its importance, I argued that there is no reason to assume that the sense of understanding is perfectly calibrated to our understanding. This makes it possible that people overestimate the detail, coherence, and depth of their understanding. Following Frank Keil, I called this phenomenon the illusion of depth of understanding. My next aim was to show that IDU is also a real possibility in scientific cognition. I argued that due to the continuity of lay and scientific cognition, the lack of explicitness in explanatory practice, the lack of clear benchmarks, the division of cognitive labor, and the future-oriented nature of scientific

cognition, there is every reason to suspect that scientists might overestimate the depth of their understanding. I further supported this possibility by presenting seven hypotheses about the possible ways in which IDU might arise in science.

What does the possibility of illusion of understanding tell us about scientific understanding? I think the main message is that we should take more care in assessing the level of our understanding. We might understand less than we think. This is not a reason for any kind of skepticism concerning the possibilities of scientific understanding, but a call for an improvement in our explanatory practices. The simple reliance on the sense of understanding is not a sufficient criterion. In order to escape the limitations of our inbuilt metacognitive apparatus, we should make the assessment of explanatory understanding more public by making our explanatory practices more explicit. In this process, the philosophy of science might be helpful, and at the end of chapter I have made some suggestions on how to recast old debates in a manner that might make them relevant to the improvement of explanatory practices and to the avoidance of IDU in science. [viii]

REFERENCES

Baker, G. P., and P. M. S. Hacker (2005), *Wittgenstein: Understanding and Meaning. Part I: Essays*. Oxford: Blackwell Publishing.

Brem, Sarah K., and Lance J. Rips (2000), "Explanation and Evidence in Informal Argument", *Cognitive Science* 24 (4): 573-604.

Clark, Andy (1997), *Being There. Putting Brain, Body, and the World Together Again*. Cambridge MA: The MIT Press.

Cleeremans, Axel, Arnaud Destrebecqz, and Maud Boyer (1998), "Implicit learning: news from the front", *Trends in Cognitive Sciences* 2 (10): 406-416.

Davis, David A., Paul E. Mazmanian, Michael Fordis, R. Van Harrison, Kevin E. Thorpe, and Laure Perrier (2006), "Accuracy of Physician Self-assessment Compared with Observed Measures of Competence", *JAMA* 269 (September 6): 1094-1102.

De Regt, Henk, and Dennis Dieks (2005), "A Contextual Approach to Scientific Understanding", *Synthese* 144: 137-170.

Donald, Merlin (1991), *Origins of the Modern Mind*. Cambridge MA: Harvard University Press.

Giere, Ronald 2002: "Scientific Cognition as Distributed Cognition", in Peter Carruthers, Stephen Stitch, and Michael Siegal (eds.), *The Cognitive Basis of Science.* Cambridge: Cambridge University Press, 285-299.

Gopnik, Alison (2000), "Explanation as Orgasm and the Drive for Causal Knowledge: The Function, Evolution, and Phenomenology of the Theory Formation System", in Frank C. Keil and Robert A. Wilson (eds.), *Explanation and Cognition.* Cambridge MA: The MIT Press, 299-324.

Grimm, Stephen R. (this volume), "Reliability and the Sense of Understanding".

Hutchins, Edwin (1995), *Cognition in the Wild*. Cambridge MA: The MIT Press.

Juslin, Peter, Anders Winman, and Henrik Olsson (2000), "Naïve Empiricism and Dogmatism in Confidence Research: A Critical Examination of the Hard-Easy Effect", *Psychological Review* 107 (2): 384-396.

Keil, Frank C. (2003), "Folkscience: coarse interpretations of a complex reality", *Trends in Cognitive Sciences* 7: 368-373.

Koehler, Derek J. (1991), "Explanation, Imagination, and Confidence in Judgment", *Psychological Bulletin* 110: 499-519.

Koriat, Asher (2000), "The Feeling of Knowing: Some Metatheoretical Implications for Consciousness and Control", *Consciousness and Cognition* 9: 149-171.

Lin, Lin-Miao, and Karen M. Zabrucky (1998), "Calibration of Comprehension: Research and Implications for Education and Instruction", *Contemporary Educational Psychology* 23: 345-391.

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

Lipton, Peter (this volume), "Understanding without Explanation".

Mills, Candice M. and Frank C. Keil (2004), "Knowing the limits of one's understanding: The development of an awareness of an illusion of explanatory depth", *Journal of Experimental Child Psychology* 87: 1-32.

Rips, Lance J. (2002) "Circular reasoning", *Cognitive Science* 26: 767-795.

Rozenblit, Leonid, and Frank C. Keil (2002) "The misunderstood limits of folk science: an illusion of explanatory depth", *Cognitive Science* 26: 521-562.

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

Schwitzgebel, Eric (1999), "Children's Theories and the Drive to Explain", *Science & Education* 8: 457-488.

Stanford, Kyle (2006), *Exceeding Our Grasp: Science, History, and the Problem of Unconceived Alternatives*. Oxford: Oxford University Press.

Trout, J. D. (2002), "Scientific Explanation and the Sense of Understanding", *Philosophy of Science* 69: 212-233.

Waskan, Jonathan A. (2006), *Models and Cognition*. Cambridge MA: The MIT Press.

Wittgenstein, Ludwig (1953), *Philosophical Investigations*. Oxford: Basil Blackwell.

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

Ylikoski, Petri (2005), "The Third Dogma Revisited", *Foundations of Science* 10, 395-419.

Ylikoski, Petri (2007), "The Idea of Contrastive *Explanandum"*, in Johannes Persson and Petri Ylikoski (eds.), *Rethinking Explanation.* Dordrecht: Springer, 27-42.

Ylikoski, Petri, and Jaakko Kuorikoski (2007), "Dissecting Explanatory Power", unpublished manuscript.

---

[i] Knowledge cannot be made completely explicit; even scientific knowledge is based on the foundation of some basic skills. For similar reasons, theoretical understanding is often not easily transformable into a more practical form of understanding. The acquisition of theoretical understanding might not be accompanied by the relevant practical knowledge and skills.

[ii] These criteria of understanding raise many questions. How are they related to each other? Are the different measures of amount of understanding comparable to each other? Could these criteria be different in different cultures or historical periods? How does one spell out the crucial notion of 'favorable circumstance' employed so many times in the above discussion? I do not pretend to be able to answer all of these questions. Here my modest aim is to describe what kind of an ability understanding is by describing some criteria according to which it is usually attributed to people.

[iii] The psychological research (Cleeremans et al. 1998) on implicit learning testifies to the plausibility of acquiring understanding without a feeling of understanding. Although the boundary between learning with and without awareness is tricky to draw, the research at least shows that one can learn without the *metacognitive* feeling that one is acquiring new knowledge.

[iv] Kyle Stanford 2006 draws from the history of science and argues it is always possible that there are equally well-confirmed and scientifically serious alternatives to current best theories that are just not conceived.

[v] A couple of psychological factors might make the evaluation of the future potential of theories even more difficult. First, the assessment of the fruitfulness of one's pet theory is probably often tainted by *wishful thinking*. From a motivational point of view this source of optimism might be instrumental in bringing about bold hypotheses, but it is not helpful in increasing the reliability of the scientist as a source of explanatory evaluation. The other factor is *the egocentric bias* in evaluation. If one is tempted to be too optimistic about one's own theories, the contrary might hold for the competing accounts. They tend to be seen as less promising. One way to bias the evaluation is to evaluate one's own theory by its future promise, and the competition by its achievements thus far. (There are various ways to rationalize this asymmetry to oneself in such a manner that the imbalance seems only fair.) My claim is not that scientists are especially disposed to be influenced by wishful thinking and egocentric bias, I am claiming rather that these factors are difficult to keep in check when people are trying to evaluate future intellectual products. The current practice of evaluating the contributions to understanding in an intuitive rather than explicit manner only exacerbates the difficulties in controlling the influence of these factors.

[vi] A similar point can be made about non-causal explanations, like constitutive explanation. In this case people think that findings about the parts of a system contribute automatically to the understanding of the capacities of the system.

[vii] The accumulation of irrelevant details makes the explanation worse, as its recipient might not see which pieces of information are relevant for understanding and which are not. But this has at least one positive consequence: at least the recipient is not prone to IDU as she fails to get the point.

[viii] I would like to thank the editors of this volume and Tarja Knuuttila, Tomi Kokkonen, Jaakko Kuorikoski, Aki Lehtinen, Uskali Mäki, Päivi Oinas and Samuli Pöyhönen for their useful comments.