View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Helsingin yliopiston digitaalinen arkisto

Anna-Mari Rusanen

ON EXPLAINING COGNITIVE PHENOMENA:

The Limits of Mechanistic Explanation

Academic dissertation

to be publicly discussed, by due permission of the Faculty of Arts
at the University of Helsinki in auditorium XV (University Main Building,
Fabianinkatu 33) on the 5th of December, 2014 at 12 o'clock.

ISBN 978-951-51-0361-1 (paperback)

ISBN 978-951-51-0362-8 (PDF)

http://ethesis.helsinki.fi

Unigrafia

Helsinki 2014

Abstract

One task of cognitive science is to explain the information processing capacities of cognitive systems, and to provide a scientific account of how cognitive systems produce the adaptive and systematic intelligent behavior that they do. However, there are several disputes and controversies among cognitive scientists about almost every aspect of the question of how to fulfill this task. Some of these disputes originate from the fundamental issue of how to explain cognitive phenomena.

In recent years, a growing number of philosophers have proposed that explanations of cognitive phenomena could be seen as instances of mechanistic explanation. In this dissertation, my aim is to examine to what extent the mechanistic account of explanation can be applied to explanations of complex cognitive phenomena, such as conceptual change.

The dissertation is composed of five related research articles, which explore different aspects of mechanistic explanations. The first two articles explore the question, whether explanations of cognitive phenomena are mechanistic in the standard sense. The third and fourth articles focus on two widely shared assumptions concerning the mechanistic account of explanatory models: that (i) explanatory models represent, describe, correspond to or are similar to mechanisms in the world and that (ii) in order to be explanatory a model must represent the relevant causal or constitutive organization of a mechanism in the world. Finally, in the fifth article a sketch of a mechanistic explanation of conceptual change is outlined.

The main conclusions of this dissertation can be summarized as four distinct, but related claims: (i) I argue that the standard mechanistic account of explanation can be applied to such cognitive explanations which track dependencies at the performance level. Those explanations refer to mechanisms which sustain or perform cognitive activity. However, (ii) if mechanistic explanations are extended to cover so-called computational or competence level explanations as well, a more liberal interpretation of the term mechanism may be needed (Rusanen & Lappi 2007; Lappi & Rusanen 2011). Moreover (iii) it is also argued that computational or competence level explanations are genuinely explanatory, and that they are more than mere descriptions of computational tasks. Rather than describing the causal basis of certain performances of the target system, or how that system can have certain capacities or competences, they explain why and how certain principles govern the possible behavior or processes of the target system. Finally, (iv) I propose that the information semantic account of representational character of scientific models can offer a naturalist account of how models depict can depict their targets, and offer also an objective account of how explanatory models can depict the relevant properties of their target systems (Rusanen & Lappi 2012; Rusanen under review).

Contents

Preface

Dissertations are typical examples of conceptual change. When doctoral students start to write their thesis, they are full of - more or less - bad intuitive ideas. During the dissertation process, the bad ideas are slowly and painfully replaced with slightly better ones. Finally, *if* the learning process is successful, one understands how deeply confused her original ideas were, and she wishes she could start it all over again.

This personal episode of conceptual change would not have been possible without the support of a group of amazing people. Firstly, I have always received help, support and advice from both of my supervisors, Prof. Petri Ylikoski and Doc. Matti Sintonen. Petri´s intellectual curiosity and ability to move across various disciplines have impacted profoundly on my attitude towards philosophical inquiry. Moreover, he has always emphasized the collective nature of inquiry both in theory and in practice. This, I believe, has created a unique environment for his doctoral students. Moreover, in addition to being a thoughtful supervisor, Petri has given me complete freedom to make my own choices. So, I have had the liberty to make my own mistakes and learn from them. Of equal importance, Matti has always supported me, when I have needed help, guidance or advice.

Secondly, I am deeply grateful to my co-author and dear friend, Dr. Otto Lappi. The first three papers of this dissertation are results of our common intellectual journey. It has been a long and winding road, but I guess we always kept our eyes on it. Thank you, Otto, for your everlasting friendship, enthusiasm, critical attitude and passion for details - they have always pushed me to go further. Moreover, without your in-depth knowledge of cognitive sciences our papers would have been far less than they turned out to be.

Thirdly, I have had the opportunity to be surrounded by active academic research groups. A very special thanks goes out to Academy Professor Uskali Mäki, whose demand for precise argumentation, conceptual clarity and exact philosophical analysis have influenced the whole group. I would also like to thank Doc. Tarja Knuuttila for her intellectual, social and practical support. I have especially enjoyed our long discussions in various airplanes. Moreover, I am deeply indebted to many other colleagues in POS, TINT and COE including Tomi Kokkonen, Inkeri Koskinen, Jaakko Kuorikoski, Aki

Lehtinen, Caterina Marchionni, Pekka Mäkelä, Samuli Pöyhönen, Jani Raerinne and Päivi Seppälä. In addition to the philosophers, I wish to thank Dr. Ismo Koponen and his research group at the Department of Physics. During the CONCHAMO project I had a wonderful opportunity to communicate with Ismo, Terhi Mäntylä, Maija Nousiainen, Tommi Kokkonen and Anu Saari ("Tommi & Annika"), whose insights on conceptual change, science learning and the nature of scientific inquiry have impacted on me deeply. Moreover, I greatly enjoyed discussions with Prof. Timo Honkela, and many other participants in our series of workshops.

This personal learning trajectory also reflects the fact that I was fortunate to be a graduate student in Cognitive Science Unit. There was, and still is, something magical in the unit, and it did, and still does, offer intellectually liberal, ambitious and genuinely multidisciplinary environment for students and researchers. Without that liberal spirit, I would have not started to write this thesis. For this reason, I want thank all the students I used to spend time with, but especially my Fodorian camaraderie Mikko Määttä.

I would also like to thank Prof. Gualtiero Piccinini for his support during these years. He has given me invaluable feedback on my papers, and his views on computational explanation have never stopped to inspire me. Moreover, I thank Prof. Paul Thagard and Prof. Oron Shagrir for the pre-examination of this dissertation, and (Sir) Henri Kauhanen for careful language revision (and demand for precision!) of this dissertation. In addition, I want to thank Carl Craver and Paavo Pylkkänen for commenting on many of my unfinished ideas.

I also thank my colleagues and friends at EAL. Especially the Spice Girls - Aino, Ellu and Laura O. - deserve their thanks. I also want to thank all of my students both in EAL and at the University of Helsinki for so many inspiring moments in the various classrooms.

Finally, I would like to thank my friends, especially Katja V., Mikke & Jemmi, my brother Heikki and his family - Nina, Iris & Atte - and my mother for their support during these years. Special thanks go to Eija and Tapio, who kindly have taken care of Ester, whenever we have needed their help. However, I am most grateful to Heikki F. and Ester just for being there.

I dedicate this dissertation to my aunt Kaarina.

List of original publications

This dissertation consists of the following publications

I. The Limits of Mechanistic Explanation in Neurocognitive Sciences

Anna-Mari Rusanen and Otto Lappi
In Vosniadou, S., Kayser, D. & A. Protopapas, (eds) *Proceedings of the European Cognitive Science Conference 2007*. Howe: Lawrence Erlbaum Associates. 2007: 284-289.

II. Turing Machines and Causal Mechanisms in Cognitive Sciences

Otto Lappi and Anna-Mari Rusanen
In P. McKay Illari, F. Russo & J. Williamson, (eds.) *Causality in the Sciences*. Oxford: Oxford University Press. 2011: 224-239

III. An Information Semantic Account of Scientific Models

Anna-Mari Rusanen and Otto Lappi
In H. De Regt, S. Hartmann, & S. Okasha, (eds) *EPSA Philosophy of Science: Amsterdam 2009*, Dordrecht, Springer, 2012: p.315-328.

IV. On Relevance

Anna-Mari Rusanen
Submitted to *European Journal for Philosophy of Science*.

V. Towards to An Explanation for Conceptual Change: A Mechanistic Alternative

Anna-Mari Rusanen
Science & Education, 23:7, (2014), 1413-1425
http://dx.doi.org/10.1007/s11191-013-9656-8)

Part I: Introductory Essay

PART I: Introductory Essay

1.1. Introduction. On Explaining Cognitive Phenomena.

It is a fundamental task of cognitive science to define and to explain the information processing capacities of cognitive systems, and to provide a scientific account of how cognitive systems produce the adaptive and systematic intelligent behavior that they do. As is well known, there are several disputes and controversies among cognitive scientists about almost every aspect of the question of how to fulfill this task. Some of these disputes originate from the fundamental issue of *how to explain* cognitive phenomena. Should we think of cognitive phenomena as subject to general, universal, law-like regularities? Or should we explain cognitive phenomena, such as computing a function, by decomposing them into sets of simpler functions and seeking how these are executed in cognitive systems?

In the history of cognitive science and psychology, examples may be found of the view according to which cognitive phenomena can be explained by appealing to general, universal and law-like regularities. For instance, German psychophysicists studied empirically the existence of law-like dependencies in the process of sensory transduction. In the domain of learning theory, behaviourists explored the possibility of formulating species-independent laws of learning (conditioning) as relationships between stimuli and responses. In recent years, some theories of associative learning can be seen as attempts to establish general and potentially universal accounts of learning (Chater & Brown 2008). Moreover, computational or probabilistic research on how people detect regularities has been characterized as a search for "universal laws of generalization" (Chater & Vitanyi 2003; Chater & Brown 2008).

This tendency to explain cognitive phenomena as subject to universal law-like regularities may reflect the idea that physics is a model for all sciences and sets the standards of explanation for other fields of science as well. However, many have emphasized that rather than adopting the explanatory strategy of the hard natural sciences, the cognitive sciences should develop such explanatory practices that align better with their actual research heuristics and explanatory purposes[1].

---

1 See, for instance Cummins 1983, 2000; Bechtel & Wright 2005; Bechtel 2008.

Namely, ever since the cognitive revolution of the 1960´s, one of the most widely used research heuristic in cognitive science has been that of "reverse engineering". In reverse engineering[2], one begins with a complex phenomenon, and then it is showed how that phenomenon can be produced by a set of less capable sub-systems. The behavior of the sub-systems can often be explained in turn by postulating the behavior of various subsystems, and so on. When this heuristic was applied to cognitive systems, it became natural to think of them as collections of highly complex computational devices, which can be studied by decomposing complex information processing tasks, such as problem solving (Newell 1980) or object vision (Marr 1982), into simpler information processing tasks. It then became possible to investigate how these simpler information processing tasks are implemented in various platforms, such as in human brains or computers[3].

It is common to think that when this heuristic is applied, the result is not typically a universal law-like generalization (Cummins 2000). Instead, the outcome of this kind of process is typically a specific model of how certain cognitive processes sustain, produce or execute the phenomenon to be explained. Recently, a growing number of philosophers and cognitive scientists have proposed that these models may provide explanations, which can be seen as instances of *mechanistic explanation* (for example, Bechtel & Wright 2005; Bechtel 2008; Piccinini 2004; Sun 2008; Kaplan 2011; Milkowski 2013).

In this dissertation, my aim is to examine to what extent this *mechanistic account of explanation* can be applied to explanations of complex cognitive phenomena, such as conceptual change. The dissertation is composed of five related research articles, which explore different aspects of mechanistic explanations. These articles can be divided into three groups. The first group explores the question, to what extent explanations of cognitive phenomena are mechanistic in the standard sense. The second group focuses on the issue of explanatory models. Finally, the third one concentrates on the question, whether conceptual change can be explained mechanistically.

---

2 Originally this heuristic was borrowed from computer science. Rather than seeking universal laws, computer scientists design solutions for specific computational problems by decomposing complex problems into their component problems. See Cummins 1983, 2000.
3 Cummins 1983, 2000.

These themes arise from the theoretical background that will be discussed in this introductory essay. The essay is organized as follows: Chapter 2 gives a general overview of the mechanistic account of explanation. In Chapter 3 the focus is on the question, to what extent the mechanistic account of explanation can be extended to cover cognitive explanations. I will first go through the argument I and Otto Lappi put forward in a pair of papers (Rusanen & Lappi 2007; Lappi & Rusanen 2011). At the end of the chapter, I will briefly present some main points of the criticism that these papers raised (Piccinini 2011; Kaplan 2011; Milkowski 2012, 2013) and reply to that criticism.

The Chapters 4 and 5 focus on the roles that models and modeling play in mechanistic explanation. In those chapters, I will examine the implications of two widely shared assumptions concerning the mechanistic account of explanatory models: That (i) explanatory models *represent, describe, correspond to* or *are similar to* mechanisms in the world and that (ii) in order to be explanatory a model must represent the *relevant* causal or constitutive organization of a mechanism in the world. As is well known, there has been a tense and vivid debate concerning both these assumptions in the philosophical literature on modeling, and I will summarize this debate briefly in Chapters 4 and 5.

Chapter 6 presents the main conclusions defended in this thesis, and discusses some implications of them in a broader framework. Finally, Chapter 7 provides an overview of the original articles.

1.2. Background: The Framework of Mechanistic Explanation

The mechanistic account of explanation was originally developed to offer an account of explanation that would stem from the explanatory practices of biochemistry and molecular genetics. Many of the traditional philosophical ideas about scientific explanations – such as the conception of laws as the principal explanatory tools in science – have been most applicable to domains of physics. In the 1970s and 1980s a group of philosophers, such as Bechtel and colleagues started to focus their attention on the biological sciences. They remarked that the traditional framework did not apply all that well to this domain. One especially important observation was that in many areas of biosciences explanations of complex systems are often given by constructing specific

models of particular mechanisms, not by offering laws or law-like general universalizations.

Since then, the mechanistic account has spread everywhere. For instance, it has been proposed to be extended to cover explanations in neuroscience (Bechtel and Richardson 1993; Craver 2007), cognitive neuroscience (Revonsuo 2006; Wright and Bechtel 2007; Kaplan 2011). It has also been suggested that computational explanation in computer science and in cognitive sciences can also take the form of mechanistic explanations (Piccinini 2004, 2007; Sun 2008; Kaplan 2011). Moreover, some philosophers of social science have argued that there are also mechanistic explanations in social sciences (Ylikoski & Hedström 2010).

1.2.1. The Mechanistic Explanation in a Nutshell

According to the mechanistic account, a phenomenon is explained by giving an accurate and sufficient description i.e. a model of how hierarchical causal systems composed of component parts and their properties sustain or produce the phenomenon (Bechtel & Richardson 1993; Machamer & al. 2000; Craver 2006, 2007). Constructing an explanatory mechanistic model thus involves mapping elements of a mechanistic model to the system of interest, so that the elements of the model correspond to identifiable constituent parts with the appropriate organization and causal powers to sustain that organization (Craver 2006; Craver & Kaplan 2011). These explanatory models should specify the initial and termination conditions for the mechanism, how it behaves under various kinds of interventions, how it is integrated with its environment, and so on.

In the mechanistic literature, it is debated whether explanations should be understood in epistemic or in ontic terms. According to the ontic view, the explanans of a phenomenon is the causal mechanism. In other words, it is the mechanism itself, rather than some representation of the mechanism that does the explaining independent of scientists' epistemic goals, abilities, explanatory purposes or interests (Craver 2007). In contrast, the epistemic view emphasizes that explanation is an essentially cognitive activity, and that mechanisms in the world do not provide explanations, while representations or models of them do (Bechtel & Abrahamsen 2005). These representations or models describe what are taken to be relevant component parts and operations, the organization of the parts and operations into a system, and the means by which operations are orchestrated so as to produce the phenomenon.

In spite of these differences, both the ontic and the epistemic view share the presuppositions that descriptions of mechanisms must represent the objective relationships between things in the real world, and that the world determines whether an explanation is correct or not. Namely, on the ontic view, mechanisms can be explanatory, if and only if mechanisms are real entities which produce or sustain the phenomena they explain, and descriptions of those mechanisms must correspond the mechanisms (Craver 2006). Also the advocates of the epistemic view require that mechanistic systems are "real systems in nature", of which mechanistic explanations provide descriptions (Bechtel and Abrahamsen 2005, p 424-425). According to this view, models – or other representations – can be used to explain, if and only if they are true, correct or accurate representations of the relevant features of their real world targets.

1.2.2. Etiological, Constitutive and Contextual Explanations

It is common to distinguish "mere" causal explanations from mechanistic explanations in philosophical literature on explanation. "Mere" or "simple" causal explanations describe the causal connection between the thing that explains and the thing to be explained i.e. they describe causal dependencies between explanans and explanandum.

Philosophers have attempted to characterize the nature of explanatory dependencies in various ways. For instance, according to one of the most influential current accounts, Woodward´s interventionist or manipulationist theory (Woodward 2003), genuine explanations offer the ability to say not merely how the system in fact behaves, but to say how it would behave under a variety of circumstances or interventions (Woodward 2003). In other words, explanation is explanation of differences, and the task of the explanans is to tell what makes the difference (Woodward 2003).

In the contrastive-counterfactual account of explanation, the relationship between explanantia and explananda are seen as objective dependencies (Ylikoski & Kuorikoski 2010; Ylikoski 2013). Explanations track these objective dependencies between phenomena and relate to these dependencies in a certain way (Ylikoski and Kuorikoski 2010; Ylikoski 2013). These dependencies can be causal, constitutive or even formal (Ylikoski 2013). Thus, in explanations the interest is in questions such as "why are things one way rather than some other way?" and so on.

A mechanistic explanation is not something that is contrary to describing what these causal or constitutive dependencies are. In contrast, mechanistic explanations describe how the dependency relation produces the phenomenon to be explained. Describing mechanisms can be seen as way of helping scientists answer these "what-would-have-happened-if-the-components-of-mechanisms-or-their-organization-were-changed"-questions. As Pöyhönen nicely puts it, these more comprehensive views on explanations "can be seen as giving a systematic account of why describing mechanisms is explanatory in the first place" (Pöyhönen 2013, p. 38).

Depending on the kinds of dependencies explanations track, mechanistic explanations can be divided into different types. *Etiological explanations* are explanations in which a phenomenon or an event is explained by describing its antecedent causes (Craver 2001). In contrast, *constitutive explanations* are explanations which explain the phenomena at a higher level of mechanistic organization by describing the internal mechanisms that produce the higher level phenomena (Craver 2001)[4]. Finally, *contextual explanations* are explanations in which the explanation is given by "showing how a mechanism is organized (spatially, temporally, and actively) into the higher-level mechanism" (Craver 2001). These explanations describe how the mechanisms are related to the other entities and activities in a higher-level mechanism, and they explain what the item does as a component of a higher-level mechanism.

These etiological, constitutive and contextual descriptions of an item's activity are three different perspectives on that item's activity in a hierarchically organized mechanism, not "metaphysical" levels of nature[5]. In other words, etiological, constitutive and contextual explanations track different kinds of dependencies. As Ylikoski (2012) puts it, while etiological explanations "are in the business of explaining changes in the properties of an entity", constitutive explanations explain "capacities, which are properties of entities or systems". Moreover, contextual explanations[6] can be seen as

---

[4] According to Craver (2001 p. 63), the relationship between lower and higher mechanistic levels is a mereological part/whole relationship with the restriction that the lower-level parts are components of (and hence organized within) the higher-level mechanism. Lower-level entities (e.g., the Xs) are proper parts of higher-level entities (S), and so the Xs are typically smaller than higher level entities, and always within the spatial boundaries of S. The activities of the lower-level parts are steps or stages in the higher-level activities.

5 Craver 2001.

6 It is not completely clear that what kind of explanations contextual explanations are. I thank Petri Ylikoski for making this point.

explanations which explain how the higher level organization of a mechanism constrains the properties and activities of lower level components.

1.2.3. On Mechanisms

While the mechanists vary slightly in their precise definitions of the term "mechanism", most of them conceive of mechanisms as causal, real and spatiotemporally localized. For instance, Glennan defines a mechanism underlying a behavior as "a complex system that *produces* that behavior by the interaction of a number of parts, where the interactions between parts can be characterized by *direct, invariant, change-relating generalizations*" (Glennan 2002, p. 344, emphasis added). Bechtel and Abrahamsen define a mechanism as "a structure *performing* a function in virtue of its component parts, component operations, and their organization" and add that "the orchestrated functioning of the mechanism is responsible for one or more phenomena" (Bechtel and Abrahamsen 2005, 423, emphasis added).

For Craver and colleagues, mechanisms are collections of entities and activities, which *are organized in the production of regular changes* from start or set up conditions to finish or termination conditions (Machamer & al, 2000; Craver 2001, 2007). On this account, a mechanism is a structure performing a function, given initial and boundary conditions, in virtue of its component parts, component operations performed by those parts, and the organization of the parts into a functional whole (the "system"). Mechanisms are something more than just aggregates of their parts (Craver 2007), and this "something more" is the "active" i.e. the spatiotemporal and causal organization of the mechanisms (Craver 2001, 2007).

1.2.4. Concrete and Abstract Mechanisms

All of these characterizations of mechanisms share the assumption that mechanisms are both causal and spatiotemporally implemented. In other words, mechanisms are seen as anchored in their components, and those components occupy space and act in real time (Craver 2001, 2007). However, recent discussions on mechanisms in various fields have demonstrated that a broader notion of mechanisms may be needed (Illari & Williamson 2011).

For instance, in cognitive sciences the notion of a computing mechanism can mean two things, either a *concrete* or an *abstract* mechanism (Piccinini 2007; Rusanen & Lappi

2007, Lappi & Rusanen 2011)[7]. When the notion of a computing mechanism is interpreted as a concrete mechanism, symbols are understood as states, and symbol structures are spatiotemporal arrangements of these symbols (Piccinini 2007). Under this interpretation computation is a *process* unfolding in real time, and computing mechanisms refer to spatiotemporally implemented computing mechanisms such as (*ex hypotesi*) neural mechanisms in brains, computers and so on.

Computing mechanisms can also be seen as *abstract mechanisms* (Piccinini 2007; Rusanen & Lappi 2007; Lappi & Rusanen 2011). Namely, computation is defined mathematically in terms of symbol structures (e.g. strings of letters produced from a finite alphabet), and instructions that generate new symbol structures from old ones in accordance with a recursive function. A computation is a proof-like sequence of symbol structures related to each other by operations specified by the instructions. As Otto Lappi and I argue (Lappi & Rusanen 2011), this definition does not necessarily involve the notion of physical causation, or require that computations are spatiotemporally implemented. For this reason, computational mechanisms can be seen as abstract in a very strong - and almost platonic sense - as mechanisms which do not operate in real space and are not spatiotemporally or causally implemented.

1.3. The Limits of Mechanistic Explanation in Cognitive Sciences

In a pair of papers Otto Lappi and I argued that the basic distinction between concrete and abstract computational mechanisms is that they operate at different levels of explanation: algorithmic (performance) and computational (competence) (Rusanen & Lappi 2007; Lappi & Rusanen 2011). These "levels" correspond to the levels of explanation or understanding introduced by David Marr (1982). Marr distinguished three distinct types of analysis of information processing systems; the computational, the algorithmic and the implementation level of analysis.

A common way to distinguish these different levels is to refer to the kinds of questions they answer. In short, computational explanations answer "what" questions and "why" questions. These are questions such as "What is the goal of the computation?", "Why does this computational task create such and such computational constraints?" and

7For other similar examples in various fields of science, see Illari & Williamson 2011. Moreover, Kuorikoski (2009) discusses on abstract and concrete mechanisms in social sciences.

"Why is this algorithm or information processing mechanism appropriate for this task?". The algorithmic level answers "how" questions ("How is this computation executed in a particular cognitive system?") by describing the specific algorithms, and the implementation level describes, how neural or other implementing mechanisms produce or sustain the behavior of the system.

1.3.1. Computational and Algorithmic Explanations

Computational explanations are explanations which focus on information processing tasks of specific cognitive capacities, such as perception, reasoning, decision making and so on. These tasks can be specified in terms of information processing as mappings from one kind of information to another.

Computational explanations track formal dependencies between certain kinds of information processing tasks and the information processing requirements of certain kinds of tasks. In a sense, computational explanations are perhaps explanations which explain why and how certain principles govern the possible behavior or processes of a system, rather than explanations of how certain mechanisms cause the behavior or processes.

While computational explanations track dependencies at the level of cognitive competences, algorithmic explanations track dependencies at the level of cognitive performances. Those dependencies, and the states of the system and the trajectories they follow, are causal (Piccinini 2007). In the case of neurocognitive sciences these performances can be found at the functional (symbolic or connectionist) or neurological (systemic, cellular, molecular) levels. These levels give a description of the concrete mechanisms that fulfill a particular cognitive task. These explanations can, for instance, explain how one ends up from one representation (which makes explicit a piece of input-information) to another (for output-information).

Thus, the basic difference between computational and algorithmic explanations is that they track different kinds of dependencies. While computational explanations track formal dependencies, algorithmic explanations track causal dependencies. Moreover, because computational, algorithmic and implementation level explanations track different kinds of dependencies, computational explanations can be considered largely autonomous with respect to the algorithmic or implementation levels below, because the

level where the cognitive capacities in question are specified is not logically dependent on the ways the causal mechanisms sustain them (Marr 1982; Shapiro, 1997). Thus the computational problems at the highest level may be formulated independently of assumptions about the algorithmic or neural mechanisms which perform the computation (Marr 1982; see also Shapiro 1997; Shagrir 2001). Generally, given a computational task, the realization can be based on various different representations and various algorithms might be appropriate for any given representation (Shagrir, 2001). In other words, these algorithms may then be multiply realized in terms of physical implementation.

1.3.2. Are Computational Explanations Mechanistic?

Piccinini proposed in a series of papers (2004, 2006a, 2006b) that computational explanations are (or should be) mechanistic explanations. Piccinini suggested that computing mechanisms can be analyzed in terms of their component parts, their functions, and their organization. A computational explanation is then "a mechanistic explanation that characterizes the inputs, outputs, and sometimes internal states of a mechanism as strings of symbols, and it provides a rule, defined over the inputs (and possibly the internal states), for generating the outputs" (Piccinini, 2006b).

In our 2007 paper (Rusanen & Lappi 2007) we argued that this kind of mechanistic account of "computational" explanations can be applied to computational explanations, if computational explanations track the dependencies at the algorithmic or performance level i.e. at the level of the mechanisms which sustain or produce the cognitive activity causally.

However, we had certain reservations about whether the mechanistic strategy Piccinini held could be extended to cover computational explanations at the competence level. Our argument was based on two claims. The first one was that computational explanations are not only bottom-up, but also top-down. The second one was that if the notion of mechanism employed in mechanistic explanation refers to concrete mechanisms only, then it cannot be applied to computational explanations at the level of cognitive competences.

1.3.3. The Problem of Direction of Interlevel Explanations

In our two papers, we took it for granted that when Piccinini discusses computational explanations, he does not view them as etiological explanations, but rather as constitutive explanations. However, in standard accounts constitutive mechanistic explanations are characterized in such a way that they seem always to be bottom-up explanations. For constitutive explanations are described as explanations in which phenomena at a higher level of hierarchical mechanistic organization are explained by their lower-level constitutive causal mechanisms but not vice versa (Craver 2001, 2006; Machamer & al, 2000). For example Craver (2001, p. 70, emphasis added) notes that "Constitutive explanations are inward and downward looking, looking within the boundaries of X to determine *the lower level mechanisms* by which it can Φ. The explanandum… is the Φ-ing of an X, and the explanans is a description of the organized σ-ing (activities) of Ps *(still lower level mechanisms)."*

However, computational explanations of cognitive competences are inter-level explanations, which also proceed top-down. Namely, computational explanations are also used to explain the behavior of mechanisms at the algorithmic and implementation levels. For instance, if one considers why a certain pattern of synaptic changes is such-and-such, one can answer: because it serves to store the value of x needed in order to compute y. Or, why is the wiring in this type of ganglion cell such-and-such? - Because the wiring computes, or approximates a computation of, some variable x. Or, why does this kind of learning mechanism lead to conceptual change? - Because it increases the utility of the conceptual system for a particular problem solving task.

In such explanations, the explanans is at the computational level, and the explananda are at the algorithmic or performance levels. In other words, the explanations start from the upper level and end up on the lower levels. But constitutive mechanistic explanations are always bottom up. Thus, we concluded, computational explanations are not constitutive mechanistic explanations in the standard sense.

However, our analysis ignores the possibility that computational explanations are *contextual* rather than *constitutive* mechanistic explanations. The contextual explanations explain how the "higher-level" mechanism constrains what a lower level mechanism does, and obviously, one computational mechanism can be a component of a larger computational system, while the latter serves as the contextual level for the

former. It may be that this is what Bechtel has in mind, when he remarks that "since computational explanations address what mechanisms are doing they focus on mechanisms "in context"" (Bechtel 2008, p. 26).

If this were right, our argument would fail. Namely, if computational-level explanations were contextual explanations, and if contextual explanation is a subspecies of standard mechanistic explanations, then computational level explanations would be a subspecies of mechanistic explanations.

However, it is still an open question, to what extent computational explanations - as we understand them - are contextual explanations *in the standard mechanistic sense*. For instance, Craver (2001, p.70) characterizes contextual explanations as explanations, which "refer to *components outside* of X" and are "upward looking because they *contextualize X within a higher level mechanism*". On this view, a description of how a cognitive system "behaves" in its environment, or how an organization of a system constraints the behavior of it´s components[8], is expressed in causal and spatiotemporal terms, not in terms of information processing at the level of computational competences.

In other words, contextual explanations show how an entity or activity fits into a *spatiotemporally* implemented higher-level mechanism. This kind of view conceives contextual explanations as a kind of systemic explanations, in which the uppermost level of the larger mechanism will still remain non-computational in character. The problem is that computational explanations do not refer to spatiotemporally implemented higher-level mechanisms and they do not involve spatiotemporally implemented components "outside of (spatiotemporally implemented) X". Instead, *they refer to abstract mechanisms, which are not spatiotemporally implemented*. For this reason, computational explanations are not these kinds of "systemic" contextual explanations.

1.3.4. Replies to Kaplan´s, Piccinini´s and Millkowski´s criticism

Some of our critics, such as Kaplan (2011) and Piccinini (2011), remark that our position can be seen as a typical example of "computational chauvinism", according to

[8] Kuorikoski & Ylikoski 2013.

which computational explanations of human cognitive capacities can be constructed and confirmed independently of details of their implementation in the brain.

In a sense, this accusation is a correct one. Indeed, we defend the view that computational explanations can be considered largely autonomous with respect to the algorithmic or implementation levels below, and that the computational problems of the highest level may be formulated independently of assumptions about the algorithmic or neural mechanisms which perform the computation (Marr 1982; see also Shapiro 1997; Shagrir 2001). Because the performance and competence- level computational explanations track different kinds of dependencies, these different modes of explanation are not necessarily *logically* dependent on each other. Hence, *if this is computational chauvinism, then we are computational chauvinists.*

However, Kaplan (2011) claims that while we highlight the independence of computational explanations, we forget something important Marr himself emphasized. Namely, Kaplan remarks that even if Marr emphasized that the same computation might be performed by any number of algorithms and implemented in any number of diverse hardwares, Marr´s position changes when he "addresses the key explanatory question of whether a given computational model or algorithmic description is appropriate for the specific target system under investigation" (Kaplan 2011, p.343).

While I am uncertain whether this is the key explanatory question Marr is interested in, I do agree that *for a cognitive neuroscientist* it is. But is this an argument against our position? As Kaplan himself remarks, Marr rejects "the idea that any computationally adequate algorithm (i.e., one that produces the same input-output transformation or computes the same function) is equally good as an explanation of how the computation *is performed* in that particular system" (Kaplan 2011 p.343, italics added)[9].

But then, *we are not talking about competence level explanations anymore*. When the issue is how the computation *is performed* in the particular system, such as in human brains, then the explanation is given in terms of algorithmic or neural processes, or mechanisms, if you will. Then, *naturally,* the crucial issue is what kinds of algorithms are possible for a certain kind of system, or whether the system has structural components that can sustain the information processing that the computational model

9 For a similar remark, see Piccinini & Craver 2011.

posits at the neural level[10]. If one aims to explain how our brains are able to perform some computations, then – of course – one should take the actual neural implementation and the constraints of the possible neurocognitive architecture into account as well.

But given this, these kinds of explanations are explanations at the algorithmic or performance level, not at the computational or competence level. For this reason, the remark that our brains cannot actually perform any kind of computations, or that the neural implementation poses some limitations to possible computations, is not actually an argument against the logical independence of the computational level[11].

A more problematic issue is to what extent these kinds of computational explanations are explanatory after all. For example, Milkowski argues that we see "functional explanations" as "a kind of systemic explanation that shows how a cognitive system can have some capacities" (Milkowski 2013, p. 107). However, we do not speak of "functional explanations" but computational explanations, and we do not claim that computational explanations explain *how* a cognitive system *can have some capacities*[12]. Instead, what we claim is that computational explanations explain *why* and *how* certain *principles govern* the possible behavior or processes of the system.

Although Milkowski may partially misinterpret our position, he still raises an important question concerning the explanatory character of computational explanations (Milkowski 2012, 2013). If computational explanations are characterized as explanations which answer questions such as: "What is the goal of this computation?", it may be claimed that we fail to make a distinction between task analysis and genuine explanations. Obviously, if computational explanations are mere descriptions of computational tasks, then they are not explanations at all.

However, as we see it, computational explanations are more than mere descriptions of computational tasks, because they describe formal dependencies between certain kinds of tasks and certain kinds of information processing requirements. If these formal dependencies are such that descriptions of them not only offer the ability to say how the computational layout of the system actually is, but also the ability to say how it would

10 See Piccinini & Craver 2011.
11 This may actually be an argument against the multiple realizability of computations rather than against the autonomy of computational explanations.
12 This is rather a question of constitution than a question of computational layout.

be under a variety of circumstances or interventions, they can be counted as explanatory[13]. In other words, if these descriptions answer questions such as "Why does this kind of task create this kind of constraint rather than that kind of constraint?" by tracking such formal dependencies which can explain what makes the difference, then these descriptions can be explanatory.

Obviously, computational explanations of this sort are not causal explanations. However, in the context of explanation of cognitive phenomena, it may be necessary to defend more liberal and pluralistic views of explanation, which would allow that there are also some non-causal forms of explanation. I agree with mechanists that when we are explaining how cognitive processing actually happens for example in human brains, it is a matter of causal explanation to tell how the neuronal structures sustain or produce the information processing in question. However, I still defend the view that there are other modes of explanation in cognitive sciences as well.

1.4. On Explanatory Mechanistic Models

In the previous sections, I have focused mostly on the issue of mechanistic explanation. In this section I turn to the issue of *explanatory mechanistic models*. This issue is important, for mechanists often emphasize that mechanistic explanations involve presenting a model of the mechanism(s) taken to be responsible for a given phenomenon, rather than presenting a formalized set of propositions (Bechtel & Abrahamsen 2005). Explanatory mechanistic models involve mapping the elements of the model to the system of interest, so that the elements of the model correspond to identifiable constituent parts having the appropriate organization and causal powers to sustain that organization (Craver 2006, Craver & Kaplan 2011).

1.4.1. Evaluation of Explanatory Adequacy

However, not all models that describe mechanisms are explanatory[14]. In contrast, the explanatory adequacy of a given model can be evaluated along three dimensions (Craver 2006, 2007), whereby distinctions may be made between (1) merely phenomenological models and genuinely explanatory models, (2) how-possibly and how-actually models, and (3) sketches and ideally complete models.

13This is a non-causal modification of Woodward (2003) and Ylikoski (2011).
[14] Moreover, not all models are models of mechanisms.

Phenomenological and explanatory models

The first distinction is between merely phenomenological and genuinely explanatory models. Merely phenomenological models provide only descriptions of the phenomena to be explained, though they allow one to make predictions about the system. However, phenomenological models do not provide genuine explanations, because they do not reveal the causal structure underlying the phenomenon that is the explanandum. For this reason, these descriptions and predictions do not show why the dependencies captured by the models are as the models describe them. Moreover, Craver suggests that genuinely explanatory models can be distinguished from merely phenomenological models by their ability to answer "what-if-things-had-been-different" questions (Craver 2006, p. 358). Explanatory models allow one to say how things would have been under different conditions, how the system would behave if it were manipulated, and so on.

How-possibly and how-actually models

The second distinction is between how-possibly and how-actually models. How-possibly models are 'only loosely constrained conjectures about the mechanism that produces the explanandum phenomenon' (Craver 2006, p. 361)[15]. Even though how-possibly models describe organized parts and activities and how they can produce the phenomenon to be explained, "one can have no idea if the conjectured parts exist and, if they do, whether they can engage in the activities attributed to them in the model" (Craver 2006, p. 361). How-actually models, on the other hand, "describe real components, activities, and organizational features" of the mechanism (Craver, 2007, p. 112). Between how-possibly and how-actually models lie the so-called how-plausibly models, which vary in their degree of realism (Craver 2006).

Craver (2006) seems to suggest that the difference between how-actually and how-possibly models is mainly a matter of how accurately or truthfully a model describes the system in question. However, as Weiskopf (2011) and Colombo et al. (forthcoming) remark, Craver´s distinction between how-possibly and how-actually models may actually be about the degrees of evidential support, and not about the truth or accuracy

[15] For alternative analysis of how possibly- models, see for instance Ylikoski & Aydinonat 2014.

of models (Weiskopf 2011)[16]. Namely, Craver's (2006, 2007) own reconstruction of the Hodgkin-Huxley model of the action potential seems to formulate the distinction in terms of confirmation rather than in terms of accuracy.

As Colombo et al. (forthcoming) note, Craver's argument that the Hodgkin-Huxley model is a how-possibly model relies on the fact that "Hodgkin and Huxley had no evidence favoring their model over other possible models". Along similar lines Weiskopf (2011) points out, any one of a set of how-possibly models might turn out to accurately model the system. But then, as Weiskopf (2011) emphasizes, if the distinction were one of truth or accuracy, it would "make little sense" to say that a how-possibly model can turn out to be a how-actually model. Instead, the dimension represents "something like the degree of confirmation of the claim that the model corresponds to the mechanism" (Weiskopf 2011, p. 316).

Sketches and complete models

The third distinction is between "sketches" and "ideally complete models". Depending on the amount of information it carries, a model can be a mechanism sketch, a mechanism schema or an ideally complete description of the mechanism (Craver 2006). A mechanism sketch is an incomplete model of a mechanism. It characterizes some parts, activities, and features of the mechanism's organization, but it has gaps. Mechanism schemata are abstracted models which omit some of the details of a particular mechanism. Ideally complete models of a mechanism include all of the entities, properties, activities, and organizational features that are relevant to every aspect of the phenomenon to be explained.

This third dimension is a matter of the representational accuracy and relevance of a model with respect to the explanandum[17]. The more features of the target system a model takes into account, the more detailed it is; and the more features relevant to the behavior of the target system a model describes, the more relevant the model is. So, the most representationally complete models are the ones that take into account all the features of the system modelled that are relevant to the behavior of the system.

---

[16] In Machamer et al. (2000, p. 21) the distinction is clearly formulated in epistemic terms as one related to "intelligibility" (Weiskopf 2011; Colombo et al, forthcoming).
[17] Weiskopf 2011, Colombo et al, forthcoming.

1.4.2. The Problem of Relevance

It has been widely recognized that the most – if any - of the models scientists use, do not provide ideally complete descriptions of their real world targets. Target systems are just too complicated to be studied in full fidelity, and thus all kinds of assumptions are made to reduce the complexity of a model. Hence, models are always more or less abstract, simplified and idealized descriptions of their real world target systems. However, in so far as models are explanatory (in a mechanistic sense), models should represent the *relevant* bits and pieces of the causal organization of their target systems[18].This raises the issue of relevance: How, exactly, are the relevant parts distinguished from the irrelevant parts?

As is well known, defining and characterizing relevance has turned out to be a notoriously difficult problem for most accounts of scientific explanation (Hitchcock 1995; Imbert 2013). For instance, in theories of causal explanation relevance is often defined as the property of causal factors that can cause the changes in explananda (Hitchcock 1995). However, only some parts of the causal chain or causal history are relevant for a given explanandum, and thus the challenge is to offer such an account of relevance, which could distinguish the relevant parts of causal history from the irrelevant parts.

Different attempts have been made to solve this problem. For example, Salmon focuses on appropriate causal relationship between the things that explain and the things to be explained (Salmon 1984; Hitchcock 1995). Salmon views the causal structure of the world as a nexus or a connection of causal processes intersecting and exchanging "marks" (1984), and argues that in etiological (i.e. causal) explanations the relevant antecedent causes are connected to phenomena by physical transmission, while irrelevant features or events are not connected in such a way. For instance, because there is no causal process which would connect the dog´s bark to the doorbell, a barking dog is not a relevant antecedent cause for the ring of the doorbell,

However, Salmon´s attempts to solve the problem of relevance have been critized in many ways. For example, Hitchcock argues (1995) that relevance is still a problem for

[18] Craver, too, refers to the pragmatics of modeling when he writes that "In fact, such (ideally complete) descriptions would include so many potential factors that they would be unwieldy for the purposes of prediction and control and utterly unilluminating to human beings"." (Craver 2006, p.5)

Salmon's account, because even if all the causal processes were identified, Salmon´s account fails to show exactly *why* the target phenomenon would occur in those circumstances. Moreover, in the context of mechanistic explanation, it does not suffice to offer an account of relevance that is appropriate for etiological explanations only: any account of relevance must be appropriate for constitutive (and contextual) explanations as well (Craver 2007b).

Craver himself proposes (2006, 2007b) that relevance can be understood in terms of the ability to manipulate one component by intervening on another, and that the relevant components or elements of a system can be established experimentally i.e. either (i) by manipulating the components and observing changes in the behavior of the mechanism as a whole or (ii) by manipulating the behavior of the mechanism as a whole and observing the behavior of the component parts. According to this account, the relevant factors are those which not only have an impact on how the system in fact behaves, but which also have an impact on how it would behave under a variety of circumstances or interventions (Woodward 2003).

Generally speaking, this kind of notion of relevance is ontic in the sense that the factors or dependencies between phenomena are seen as objective features of the world. In other words, they are not dependent on scientists' ways of conceptualizing or theorizing about them, and the degree of relevance is not dependent on our psychological, pragmatic or inferential practices. However, Craver (2006) seems to allow for the possibility that relevance may also involve some pragmatical dimensions. He makes the following – quite cryptic – remark (2006, p. 360):

> "Models *frequently drop details that are irrelevant in the conditions under which the model is to be used*… Which information is relevant varies from context to context… Explanations are sometimes represented as answers to questions about why something has happened or about how something works. Questions and answers, as linguistic entities, presuppose a conversational context that specifies precisely what is to be explained (by a model) and how much detail will suffice for a satisfying answer".

If I understand Craver correctly, he seems to think that the relevance (of the details of a model) is partially dependent on the pragmatical context of modeling. In other words, some elements *are selected* as relevant (or expressed on some continuum of relevance) from a number of available existing elements, because they are (or are seen) relevant for

the explanatory purpose(s). Some other elements or details are considered irrelevant, because they are not (seen) relevant for these purposes.

This kind of relevance can be called "pragmatic relevance" (Rusanen under review). Generally, this "pragmatic relevance" can be interpreted either as *intentionality based* or as *task based* relevance (Rusanen under review). According to the intentionality based account, the relevance of P for a certain task Z is understood as dependent on the model user´s (intentional) evaluation of the usability of P for certain purposes, such as explaining. This kind of *subjective, intentionality based relevance* can be characterized as follows:

> (Subjective relevance) *P is relevant to Q, if and only if P is judged (intentionally) to be relevant to Q.*

This has been quite a popular view in the philosophical literature on modelling, and many have shared the intuition that the relevant properties are those that the users of models take, interpret or even intend to be relevant (for example Giere 2004; Contessa 2011). For example, Giere (2004, italics added) writes as follows:

> "...How do scientists use models to represent aspects of the world? …One way... is by exploiting similarities between a model and that aspect of the world it is being used to represent... It is not the model that is doing the representing; it is the scientist using the model who is doing the representing. Part of being able to use a model to represent some aspect of the world is being able to *pick out the relevantly similar* features...."

According to this view, relevance is based on subjective and intentional relevance assessments, on the ways of conceptualizing the tasks, on personal expectations and even on the motivation behind the scientific research. However, this does not mean that these subjective relevance evaluations are completely individual. On the contrary, these evaluations are often influenced by socio-cultural standards of relevance assessments, the strategical and tactical decisions by groups of scientists, the division of epistemic labour, and the normative practices of the scientific community at large.

However, I doubt whether advocates of mechanistic explanation are willing to accept this view of relevance. Merely postulating or intending that a model captures the relevant causal organization does not guarantee that the model actually captures the relevant bits and pieces. For a mechanist, it is crucial that those factors that are taken to

be relevant really do correspond to the relevant aspects of the world. Explanation is not a matter of intending explanatory relationships to hold, but a matter of describing how the world really is. For this reason, I believe that Craver's cryptic remark may actually refer to the idea that *fulfilling* a certain task, such as explanation, determines the pragmatic relevance. This kind of relevance can be called *task based relevance*, and it can be characterized as follows:

> (Task Based Relevance) *P is relevant to task T, if and only if it increases the likelihood of accomplishing the purpose Z which is implied by the task*[19].

If the "usability" for a task is what determines the relevance of P, then it is still possible to think that "usability" is not dependent on our subjective ways of evaluating that usability. In short, P can be used for the purpose Z (such as explanation), if P is suitable for it (whether or not the user realizes, or even evaluates the relevance of P for it). If explanation is understood in terms of manipulation and control, then the degree of P´s relevance is dependent on its ability to depict the right kinds of causal factors.

1.4.3. Explanatory Mechanistic Models and Scientific Realism

What emerges, then, is a picture of explanatory models in which these models do not only include information, but are required to be adequate *how-actually* explanations of the relevant features of target systems. In short, they explain the phenomenon by *giving an accurate* and *sufficient description* i.e. a model *of how the relevant hierarchical causal systems* composed of component parts and their properties sustain or produce the phenomenon in the world (Bechtel & Richardson 1993; Machamer & al. 2000; Craver 2006, 2007).

In a sense, this picture of explanatory models is a typical example of a (possibly extreme) realistic attitude towards models and modeling. According to realism, scientific explanations, theories and models are (at least approximately) correct descriptions of how things stand in the real world. This realistic attitude involves a metaphysical claim and a semantic claim. The metaphysical claim is that the components and activities, of which mechanisms consist, are real and mind-independent features of the world (Bechtel and Abrahamsen 2005, p. 423-5; Machamer, Darden and

[19] There is an analogical notion of relevance in information and library sciences. See Sarasevic 2007 for an introduction.

Craver 2000; Craver 2006). The semantic claim is that our explanatory models give us descriptions of their targets, because they can correctly *represent, depict* or *describe* the causal interactions among the parts of the target mechanisms that enable the latter to produce the phenomena they do under various conditions.

Generally, in the mechanistic literature, this representational relationship -"the aboutness"- between a model and its target systems is often seen as some kind of similarity or correspondence relation. For instance Glennan characterizes the representational relationship as "*one of approximate similarity*" (Glennan, 2000) in a way, where "the behavior of the system in nature is described (to varying degrees of approximation) by the model's behavioral description and the internal structure of the system is described (again to varying degrees of approximation) by the model's mechanical description". Craver and Kaplan (2011) describe the representational relationship as "*correspondence*" in the following way "…in succesfull explanatory models… (a) the variables in the model correspond to components, activities, properties and organizational features of the target mechanism that produces, maintains, or underlies the phenomenon, and (b) the… dependencies posited among these variables in the model correspond to the… causal relations among the components of the target mechanism".

However, as already Quine (1969) pointed out, similarity is a vague notion, and correspondence is actually not much better. In order to specify these concepts, many philosophers of science have appealed various "morphisms" – iso-, partial- or homomorphisms. However, all these various morphisms, similarities and correspondences are problematic, and they have been critized on many grounds.

In the next section, I will briefly summarize some of the main points of this criticism, and then briefly describe the basic tenets of the information semantic account of scientific representations.

1.5. On the Problem of Scientific Representation

Generally, representations are things that we can think of as standing in for something. A photograph of a person stands in for that person. A model of DNA represents the structure of the DNA molecule: bits of the model stand in for bits of DNA (bases,

sugars, phosphates) and their relations stand in for the relations between the bits (e.g. base pairing).

The directionality of representations poses some formal requirements for representational relationships between the things that represent (i.e. the representations) and the things that are represented (i.e. the targets) (Cummins 1983; in the context of scientific representations, see Suaréz 2003). Firstly, a representational relation[20] is *asymmetric*: representations stand in for their targets, but targets do not stand in for the representations. For instance, a photograph of a person stands in for that person, but the person does not represent the photograph. In a similar way, if a model of DNA represents DNA, then DNA does not represent the model. Secondly, representations *are not reflexive*, since they rarely, if ever, represent themselves. A photograph of a person represents the person, but the photograph does not represent itself. In a similar way, a model of DNA represents DNA, but the model of DNA does not represent the model of DNA. Thirdly, it is often emphasized that representations are also *intransitive*: Even if the target B of a representation A were itself a representation of some further target S, yet the representation A would not represent S, but only B. For example, if a painting (B) represents a person (S), then a photograph of the painting (A) represents just the painting, not the person.

1.5.1. On Accounts of Scientific Representations

Recent discussions of scientific representations offer what may appear to be three broad and conflicting approaches. Firstly, some approaches emphasize that a scientific representation is something that bears an objective relation to the thing it represents. In what follows, I will call proponents of such approaches "*naturalists*". Secondly, there are "*intentionalist*" accounts, according to which models represent their targets in virtue of the modeler's intentions to use them as representations. Thirdly, there are accounts which urge us to adopt the "deflationary attitude" towards scientific representation. These *inferential* or *minimalist* conceptions of scientific representation reject the view that scientific representation is "an substantive objective relation" between A and B that answers "solely to the properties of A and B" (Suárez 2004).

The Minimalist Approach

[20] See Suaréz 2003.

Perhaps the most well-known advocate of the minimalist or inferentialist account is Mauricio Suárez, who in a series of papers (2002, 2003, 2004) has argued for the "deflationary attitude". According to this attitude, attempts to formulate a general account of scientific representations are doomed to fail, because they track a special relationship which does not exist. For this reason, we should not seek for "deeper" features of representation other than its surface features (Suárez 2002). According to Suárez, these "surface features" are: (1) the representational force of a source, and (2) the capacity of surrogate reasoning i.e. the capacity to draw inferences about the target from the model. Thus, according to this minimalist approach A represents B if (and only if) (i) the representational force of A points toward B, and (ii) A allows competent and informed agents to draw specific inferences regarding B (Suárez, 2002, p. 27).

However, it is unclear how minimalist Suaréz´s minimalism actually is. Namely, he defines the representational force of a model as "the capacity of a source to lead a competent and informed user to a consideration of the target", in virtue of "a relational and contextual property of the source, fixed and maintained in part by the *intended representational uses* of the source on the part of agents" (Suárez 2004, p.773, italics added). This seems to suggest that it is the intended uses, and a fortiori the intentions of users, which are – at the end of the day - doing the representational work[21]. Nonethless, it is possible to defend such an interpretation of inferentialism, which emphasize that the representational "properties" of models are reducible, or based on the inferences the models allow[22]. On this view, the inferential properties of models constitute their representational properties, and a model is a representation of a certain target (for a certain user), if and only if the user is able to perform "surrogative inferences" from a model to the target. Nevertheless, also this line of minimalistic reasoning is problematic. It seems counterintuitive to suggest that a model can be used to represent (the relevant features of) the world, *because* we can make inferences on the basis of it. As Contessa (2011) remarks, it is more intuitive to think that we can make correct inferences on the

[21] Actually, Suárez and Solé (2006, p.39) seem to admit this by writing how "representation in science is conceived as an intentional activity… The inferential conception takes it that agents' pragmatic purposes are essential in two different ways to the nature of the kind of cognitive representations one finds in science: i) as initial fixers of the representational force that points from A to B when A represents B, and ii) as defining the level of information, skill and competence required for an appropriate use of the representation, which in turn determines the inferential capacities of the source —i.e. it determines the inferences about B that can legitimately be carried out on the basis of reasoning about A.

[22] See Ylikoski & Kuorikoski, under evaluation.

basis of models, *because* models depict or refer to the real world targets in one way or another[23].

Naturalist Accounts

Naturalists suggest that there is a mind-independent, objective "representational" relationship between models and their targets. According to *similarity based* views, a model represents its target in virtue of the fact that the former is similar to the latter. However, as many have pointed out, this conception is problematic (for example, Cummins 1989; Suárez 2003). Firstly, similarity does not fulfil the requirements set for the representational relation, because the similarity relation lacks the distinctive "directionality" of representations: similarity relations are reflexive and symmetric, while the representation relation is irreflexive and asymmetric (Cummins 1989; Suaréz 2003). Secondly, similarity is a vague notion, which needs a more rigorous characterization.

Many philosophers have appealed to various "morphisms" (isomorphism, partial isomorphism, homomorphism) in order to specify – or partially in order to replace – the notion of "similarity" (van Fraassen 1980; da Costa and French 2000; French 2003 etc.). However, also these attempts have been critized in many ways. Just like similarity, the various morphisms are symmetric, reflexive and transitive relations, and for this reason they do not qualify as representations (Suárez 2003). Moreover, both the similarity and isomorphism-based accounts have trouble with the problem of representational relevance. A model typically cannot be perfectly "similar" or "isomorphic" with its target system, since target systems are too complex. For this reason, it is often, if not always necessary to reduce this complexity by making all kinds of simplifying and unrealistic assumptions. Thus, to save the representational relationship, models should be "sufficiently similar", "similar in specified aspects" or "sufficiently isomorphic" to the target in the relevant respects. However, it is quite tricky to characterize "sufficiency" and "relevance" in a precise manner – especially in naturalist terms. This may partially explain why so many philosophers have been tempted into intentionality based semantics. It is simply intuitively compelling to think that the relevant properties of

[23] As Chakravarty (2010) remarks, even Suárez himself (2003, p. 229) admits that "if A represents B, then A must hold some particular relationship to B that allows us to infer some features of B by investigating A".

models are those that the users of models take, interpret or even intend to be relevant (for example Giere 2004; also Contessa 2011).

Intentionality Based Accounts

According to the intentionality based account, models represent their targets in virtue of the modelers' intentions to use them as representations. In other words, models have their semantic properties in virtue of the intentional states of model users. For instance, Teller (2001, p.397) describes this approach nicely by writing that "I take the stand that, in principle, anything can be a model, and that what makes a thing a model is the fact that it is regarded or used as a representation of something by the model users."

The cost of this "intentionalist" or "derived intentionality" – as Searle (1992) would call it -– approach is that it will make accounts of scientific representations dependent on prior intentional characterization of the users, and on empirical facts about how scientists interpret their models. However, since empirically the intentional practices of scientists are complicated and not at all well understood, the issue becomes unnecessarily complicated. "Intending to represent scientifically" is not better understood than "representing scientifically", in fact less so. Moreover, as many have pointed out, merely postulating a representational relation to hold between a model and an intended target (or intending such a relation to hold) does not create a proper representational relation between them (Frigg 2006). As Frigg writes, "to say S is turned into a representation because a scientist intends S to represent T is a paraphrase of the problem [of giving an account of scientific representation – of explaining why or how S represents T] rather than a solution" (Frigg 2006, p. 54).

In addition, the problem of relevance arises again. To say that S is turned into a representation of the relevant features of a target system T, because a scientist intends S to represent the relevant aspects of T, is also a paraphrase of the problem of relevance rather than a good solution to it. Moreover, "intending to represent the relevant parts of a target system" is not better understood than mere intending S to represent T. In contrast, appealing to intentions – again -– just complicates the problem further. Empirically speaking, there is no widely accepted theory of relevance among cognitive scientists or psychologists. On the contrary, for the last 60 years cognitive scientists have tried to understand how human cognition is able to realize relevance, to evaluate relevance, and to use relevance in various forms of cognitive processes. But no

consensus has emerged, and explaining relevance has actually turned out to be one of the most difficult tasks for cognitive scientists[24].

1.5.2. The Plea for Objective Accounts of Semantic Relevance

Because merely postulating or intending that a model captures the relevant causal organization does not guarantee that the model actually captures the relevant bits and pieces, it may be that solving the problem of semantic relevance requires postulation of at least some sort of objective relation between a model and its target.

In some approaches isomorphism (or another morphism) is meant to offer such a substantive connection. In those accounts isomorphism regulates the way in which the (intended) model relates to the target system, and thus imposes constraints on what kinds of representation are admissible (Frigg 2006). However, isomorphism does not offer a sufficient constraint on relevance, because a prior choice of the relevant similarities is needed to get any isomorphism based account off the ground. To put this in a slightly different way, before one can inquire into the isomorphism of two structures, one must first identify the "relevant" elements and relations. This identification may be done by appealing intentions of users or by appealing to some sort of objective factors. If this identification is based on modelers' intentions, then we are back with problems of intentionality based accounts. Hence, the solution for the problem of relevance may require a more objective stance.

1.5.3. The Information Semantic Account of Models and the Problem of Relevance

In our 2012 paper I and Otto Lappi developed a completely new account of the representational character of scientific models, which we titled "the information semantic account of scientific models" (Rusanen & Lappi 2012a). According to this account, a model represents its target if and only if it carries information about its target[25].

---

[24] Carruthers 2003; Shanahan 2003.

[25] This view of models and modeling differs from the picture of modeling portrayed for instance by Weisberg (2007). According to Weisberg, modeling can be seen as the indirect theoretical investigation of a real world phenomenon using the model as a "mediator". On this view, this investigation happens in three stages: In the first stage, a theorist constructs a model. In the second, she analyzes, refines and further articulates the properties and dynamics of the model. Finally, in the third stage, she assesses the

This information connection between the parts of the model and parts of the target system is implemented in model building. It includes, for example, the data collection, model fitting, different experimental and data analysis methods and so on. This kind of process is an iterative and self-correcting process of interaction with the phenomenon, and at its best it ensures non-accidental convergence of some aspects of the world and the structure or behaviour of (some parts of) the model (Rusanen & Lappi 2012a).

In practice, this is what scientists often try to do. When scientists build a model of some target system they do a lot of systematic work to ensure that the model is connected to its target rather than something else, e.g. the behavior of the measuring device (artifacts), or nothing at all (noise). For instance, scientists debug their data gathering methods, analyse their operational criteria and data analysis methods, do parallel experiments to verify assumptions related to hypotheses concerning both the real world phenomena and the behaviour of measuring devices and they relate their model to a known background theory of the functioning of their measurement devices etc. In this way, scientists do not only manipulate the properties of the target system and record the systematic effects of these manipulations, but they also conduct a lot of parallel methodological research on the parts of the data gathering and analysis methods in order to be able to present sufficient evidence that the model is referring to elements of target system.

In our paper (2012a) we proposed that the representational character of models is then a product of, and is in part defined by, this iterative model building process, in which information about a phenomenon comes through from the empirical data. If the model building process is successful, then this information is incorporated into the model. The semantics of models is thus a result of information carrying properties of the models that emerge from the model building process, not the modelers´ intentions per se. The model building processes may well be directed by the modellers' assumptions and purposes, as well as by assumptions about mappings made by the modellers. However, the intentions or purposes, which causally direct the model building process, do not enter into the definition of the semantic relation itself. Hence, in the context of

---

relationship between the model and the world if such an assessment is appropriate. If the model is sufficiently similar to the world, then the analysis of the model is also, indirectly, an analysis of the properties of the real world phenomenon. In our account, the first stage, the construction of models, involves data gathering, data-analysis and so on.

information semantics the semantic constitution of a model is based on an objective model-world relation, not on the purposes or intentions of a scientist.

This reflects the fact that in information semantics, the semantics and the pragmatics of models are kept distinct. The semantics of a model is about how a model "represents", "depicts" or "is about" its target system. The pragmatics of models, on the other hand, concerns the context-dependent properties and features of the ways that models are used in scientific practices, including the ways that they are used to explain in certain contexts, to make predictions about the possible behavior of target systems, and analyze the structure of the target system. The semantic constitution of a model is more fundamental than the pragmatics in the following sense: If a representation A does not carry information about the target B, A is really not a representation of the target B – even if A is considered by modelers "a representation of B". On this view, unless a model carries information about its target, it fails to represent the target altogether (the modelers' intentions notwithstanding), and if a model carries information about its target the model represents this target – whether or not its user realizes it or even intends it. In information semantics it is the world, which decides, which models are representing their targets, and which are not.

The semantics also places some restrictions on the meaningful use of a model: the usefulness of models for representational purposes depends on their semantic relation to their target systems. As a semantic theory, the information theoretic account of semantics is not only descriptive, but also normative. It gives a criterion for distinguishing a "genuine" representation from arbitrary mappings: genuine information carrying representations allow us to obtain information about the intrinsic properties of target systems, arbitrary or false mappings do not.

Moreover, we suggested that at its best the iterative model building process ensures non-accidental convergence of some aspects of the world and the structure or behaviour of (some parts of) the model. On this view, the semantically "relevant" aspects of the world X are simply the parts of the world that this kind of model building process ends up tracking, and the relevant parts of the model F are the ones that end up performing this tracking – whether or not these properties are the ones the model builders intend or actually believe to be involved. Thus the ability to refer to relevant dependencies is a

result of the information carrying properties of the models that emerge from the model building process, and not of the modelers´ intentions per se.

1.5.4. Concluding remarks: The Virtues of Information Semantic Account

A number of theoretical virtues result from thinking of representational features of models in this way. Firstly, in information semantics the representational relationship is a directional relationship, and for this reason the information semantics fulfils the requirements of asymmetry and irreflexivity (Rusanen & Lappi 2012a)[26]. Secondly, it also solves the problem of circularity (Rusanen & Lappi 2012a). Thirdly, it may shed light on the issue, how to define the relevance relation between models and their targets in objective terms (Rusanen & Lappi 2012a, Rusanen under review). Finally, it offers an opportunity to define the semantics of scientific representations directly, without reference to prior users' intentions.

Of course, there are many problems left open by the information semantic account. For example, it is extremely complicated to spell out how, exactly, models carry information about their targets, and how the flow of information is implemented in a model building process. Moreover, traditionally information semantics has been mostly applied only to mental representations, and one may raise the question to what extent it is an appropriate account of scientific representations[27]. However, there are some information semantic based accounts, such as Ryder´s account (2004), in which brains are described as "model making machines", and representations are seen as model-like entities produced by the brains. If mental representations are conceived in this way, the gap between mental and scientific representations may not be as deep as it seems to be.

There is also a set of problems related to fictional models, which may pose some challenges to the information semantic account. However, as we have argued elsewhere (Rusanen & Lappi 2012b), many of these problems are strictly analogous to problems which crop up in information semantics in the philosophy of mind. These problems have been extensively discussed there since the 1980's, and there are significant recent developments.

---

[26] As such, information semantic account is not in contradiction with other naturalist accounts, such as similarity-, resemblance- or isomorphism based views. In contrast, it is possible to supplement these accounts with information semantic account.

[27] I thank Oron Shagrir for raising this interesting question.

1.6. Conclusions

The main conclusions of this dissertation can be summarized as four distinct, but related claims. The first claim concerns the applicability of mechanistic explanation to cognitive sciences, the second is about the explanatory character of computational explanations. The third claim concerns the representational character of explanatory models, and the fourth one focuses on the issue of relevance.

More precisely, in this dissertation I argue that

1) The standard mechanistic account of explanation can be applied to such cognitive explanations which track dependencies at the performance level. Those explanations refer to mechanisms which sustain or perform cognitive activity. For instance, those explanations of conceptual change that focus on individual learning trajectories can be seen as instances of mechanistic explanation (Rusanen 2013).

2) However, if mechanistic explanations are extended to cover so-called computational or competence level explanations as well, a more liberal interpretation of the term mechanism may be needed (Rusanen & Lappi 2007; Lappi & Rusanen 2011). Typically mechanists define mechanisms as concrete mechanisms, which are implemented spatiotemporally. This notion of mechanism is not compatible with the notion of abstract mechanism that is used in competence level explanations (Rusanen & Lappi 2007; Lappi & Rusanen 2011).

3) Computational or competence level explanations are genuinely explanatory, and they are more than mere descriptions of computational tasks. If explanations differ from mere phenomenological descriptions in that they allow control and manipulation of the system in question, and that they also allow us to answer various counterfactual questions about the system's behavior (Craver 2006), then computational explanations, too, can be explanatory. Namely, computational explanations describe formal dependencies between certain kinds of tasks and certain kinds of information processing requirements. If these formal dependencies are such that descriptions of them not only enable one to say what the computational layout of the system actually is, but to say how it

would be under a variety of circumstances or interventions, they can be counted as explanatory[28]

4) Obviously, computational explanations of this sort are not causal explanations. Rather than describing the causal basis of certain performances of the target system, or how that system can have certain capacities or competences, they explain why and how certain principles govern the possible behavior or processes of the target system.

5) The "information semantic account of scientific models" provides a naturalist view of how scientific models can depict their target systems (Rusanen & Lappi 2011). According to this view, a model represents its target if and only if it carries information about its target. The information connection between parts of the model and parts of the target system is implemented in the iterative and self-correcting process of model building. At its best it ensures non-accidental convergence of some aspects of the world and the structure or behaviour of (some parts of) the model (Rusanen and Lappi 2012).

6) In information semantics, the representational relationship is a directional relationship. For this reason, the information semantic account can meet the requirements of asymmetry and irreflexivity (Cummins 1983; in the context of scientific representation, see Suaréz 2003). Moreover, the information semantic account can offer a possible naturalistic solution to the problem of relevance that has posed difficulties for most naturalistic accounts of scientific representations. On this view, the semantically "relevant" aspects of the world X are simply those parts of the world that the iterative and self-correcting model building process ends up tracking, and the relevant parts of the model F are the ones that end up performing this tracking. The model building process may well be directed by the modellers' intentions, as well as by assumptions about mappings made by the modellers. However, the intentions or purposes, which causally direct the model building process, do not enter into the definition of the relevance relation itself. The ability to refer to relevant dependencies is a result

[28] This is a non-causal modification of Woodward 2003 and Ylikoski 2011.

of information carrying properties of the models that emerge from the model building process, not of the modelers´ intentions.

7) The information semantic account may also offer an objective account of how explanatory models can depict the relevant properties of their target systems (Rusanen under review). Explanatory models track or detect the causal, constitutive or formal dependencies of their target systems. According to the information semantic account, if a model building process ends up tracking these dependencies, then the relevant parts of the model F are the ones that end up performing this tracking.

1.6.2. Final Remarks. The Plea for Explanatory Pluralism.

In this dissertation, I have argued that some explanations of cognitive phenomena can be subsumed under the banner of "mechanistic explanation". Typically those explanations are neurocognitive explanations of how certain neurocognitive mechanisms produce or sustain certain cognitive phenomena, but also some psychological explanations can be seen as instances of mechanistic explanations (Rusanen 2013). Moreover, if a more liberal interpretation for the term mechanism is allowed, then *some* computational or competence level explanations may also qualify as mechanistic explanations (Rusanen & Lappi 2007; Lappi & Rusanen 2011).

Nevertheless, I doubt whether mechanistic explanation can be extended to cover *all* cognitive explanations. There are several reasons for this plea for explanatory pluralism: Firstly, it is not clear whether all cognitive systems or cognitive phenomena can be captured mechanistically. Mechanistic explanations require that the system can be decomposed i.e. analyzed into a set of possible component operations that would be sufficient to produce or sustain the phenomenon in question (Bechtel & Richardson 1993). Typically a mechanism built in such a manner will work in a sequential order, so that the contributions of each component can be examined separately (Bechtel & Richardson 1993).

However, in cognitive sciences there are examples of systems – such as certain neural nets – which are not organized in such a manner. As Bechtel and colleagues remark, the behavior of these kinds of systems cannot be explained by decomposing the systems

into subsystems, because the parts of the networks do not perform any activities individually that could be characterized in terms of what the whole network does (Bechtel & Richardson 1993; Bechtel 2011, 2012). Hence, it is an open question to what extent the behavior of these kinds of systems can be explained mechanistically. At the very least, it will require adopting a framework of mechanistic explanation different from the one that assumes sequential operation of decomposable parts (Bechtel 2011, 2012; Bechtel & Abrahamsen 2011).

Moreover, Von Eckardt and Poland (2004) raise the question to what extent the mechanistic account is appropriate for those explanations which involve appeal to mental representations or to the normative features of certain psychopathological phenomena. Although I find Von Eckardt and Poland´s argumentation slightly misguided, I still think that it is important to consider the normative aspects of cognitive phenomena. Cognitive systems are, after all, adaptive systems which have a tendency to seek "optimal", "rational" or "best possible" solutions to the information processing problems that they face[29]. Because of this, cognitive processes are not only goal-directed, but also normative. It is not clear how well this normative aspect of cognitive systems can be captured by mechanistic explanations.

Thirdly, some philosophers have paid attention to the fact that there are examples of explanatory computational models in cognitive sciences which focus on the flow of information through a system rather than the mechanisms that underlie the information prosessing (Shagrir 2006, 2010). Along similar lines, Weiskopf (2011) argues that there is a set of "functional" models of psychological capacities which are both explanatory and non-mechanistic.

Finally, in recent years cognitive scientists have raised the possibility that there are some universal, law-like principles of cognition, such as the "principle of simplicity", "universal law of generalization" or the "principle of scale-variance" (Chater & Brown 2008; Chater & Vitanyi 2003). Chater and colleagues (ibid.) argue that it is possible to explain many cognitive phenomena, such as certain forms of linguistic patterns, or certain types of inductive generalizations, by combining these principles. These

[29] Of course, often these solutions are not ideally optimal or ideally best possible ones. They are typically (if not always) results of different kinds of performance limitations, trade-offs between different aspects of solutions and so on.

explanations are “principle based” rather than mechanistic explanations. Moreover, Chater and colleagues seem to suggest the mechanistic models of these phenomena may actually be derived from these general principles, and explanations that appeal to these general principles provide “deeper” explanations than the mechanistic explanations (Chater & Brown 2008). It is possible, that many of the so called computational level explanations turn out to be instances of these principle-based explanations rather than instances of mechanistic explanations.

In sum, taken together these diverse claims seem to imply that there is not a single, unified mode of explanation in cognitive sciences. Instead, they seem to suggest that cognitive sciences are examples of those sciences which utilize several different modes of explanation, only some of which can be subsumed under the mechanistic account of explanation. Obviously, mechanistic explanation is a powerful framework for explaining the behavior of complex systems, and it has demonstrated its usefulness in many scientific domains. Also, many successful theories and explanations in cognitive sciences are due to this mechanistic approach. However, this does not imply that it would be the *only* way to explain complex cognitive phenomena.

1.7. Overview of Original Articles

This dissertation is composed of five articles. The first two articles explore the question, to what extent explanations of cognitive phenomena are mechanistic in the standard sense. The third article focuses on the issue of the representational character of scientific models. The subject of the fourth article is the notion of relevance. Finally, in the fifth article a sketch of a mechanistic explanation of conceptual change is outlined.

1.7.1. The Limits of Mechanistic Explanation in Cognitive Neurosciences

Piccinini proposed in a series of papers (2004, 2006a, 2006b, 2006c) that computational explanations are mechanistic explanations. He suggested that computing mechanisms can be analyzed in terms of their component parts, their functions, and their organization. A computational explanation is then "a mechanistic explanation that characterizes the inputs, outputs, and sometimes internal states of a mechanism as strings of symbols, and it provides a rule, defined over the inputs (and possibly the internal states), for generating the outputs" (Piccinini, 2006b).

In "The Limits of Mechanistic Explanation in Cognitive Neurosciences" I and Otto Lappi argue that this kind of mechanistic account of "computational" explanations can be applied to such explanations which track causal dependencies at the level of cognitive performances. In those explanations, the explanantia can be given in terms of causal mechanisms. These explanations correspond to explanations which Marr called "algorithmic" or "implementation-level" explanations.

However, we found it difficult to assume that the standard mechanistic account of explanation could be extended to cover explanations which track dependencies at the level of cognitive competences i.e. the computational explanations in Marr´s typology.

Our argument was based on two claims. Firstly, we claimed that there are examples of inter-level computational explanations, which can proceed top-down. In those cases, computational explanations explain also the behavior of mechanisms at the algorithmic and implementation levels. In such explanations, the explanans is at the computational level, and the explananda are at the algorithmic or performance levels. In other words, the explanations start from the upper level and end up on the lower levels. If this is right, then this mode of explanation does not correspond to the standard picture of constitutive

mechanistic explanation. In standard accounts (constitutive) mechanistic explanations are characterized in such a way that they seem always to be bottom-up explanations. In those explanations, phenomena at a higher level of hierarchical mechanistic organization are explained by their lower-level constitutive causal mechanisms but not vice versa (Craver 2001, 2006; Machamer & al, 2000).

Secondly, we remarked that if mechanistic explanation requires that both the explananda and the explanantia are causal, then explanations which appeal to the abstract competence level just cannot serve as a source of standard mechanistic explanations. The computational information structure of a cognitive system is not characterized in terms of causal organization. In contrast, we claimed, it involves abstract mechanisms, which are not causally, but logically governing the behavior of the mechanisms at the lower levels.

1.7.2. Turing Machines and Causal Mechanisms

In "Turing Machines and Causal Mechanisms" we developed the distinction between abstract and concrete computational mechanisms further. In cognitive sciences the notion of a computing mechanism can mean two things, either a concrete or an abstract mechanism (Piccinini 2007). When the notion of a computing mechanism is interpreted as a concrete mechanism, computation is seen as a process unfolding in real time, and computing mechanisms refer to spatiotemporally implemented computing mechanisms such as neural mechanisms in brains and so on.

Computing mechanisms can also be seen as abstract mechanisms (Piccinini 2007). In "Turing Machines" we argue that these abstract mechanisms do not necessarily involve the notion of physical causation, or require that computations are spatiotemporally implemented. On this view, computational mechanisms can be abstract in a very strong sense: They can be viewed as mechanisms which do not operate in real space, and they are not spatiotemporally implemented causal mechanisms.

1.7.3. Information Semantic Account of Scientific Models

In "Information Semantic Account of Scientific Models", a novel account of scientific models is presented. In this information semantic account, models are understood as information carrying artifacts. According to this proposal, models "represent", "depict"

or “are about” their real world target systems, if they carry information about relevant parts of the target systems.

The information connection between the parts of the model and parts of the target system is implemented in model building. It includes, for example, the data collection, model fitting, different experimental and data analysis methods and so on. This kind of process is an iterative and self-correcting process of interaction with the phenomenon, and at its best it ensures non-accidental convergence of some aspects of the world and the structure or behaviour of (some parts of) the model.

The representational character of models is then a product of, and is in part defined by, this iterative model building process, in which information about a phenomenon comes through from the empirical data. If the model building process is successful, then this information is incorporated into the model. The semantics of models is thus a result of information carrying properties of the models that emerge from the model building process, and the semantic constitution of a model is based on an objective model-world relation.

1.7.4. On Relevance

“On Relevance” proposes a short characterization of the general notion of relevance. According to it, relevance can be viewed as “a relation between a P (or a number of Ps) and a Q (or a number of Qs) along some property R (or a number of Rs), such as utility, topicality, information and causality etc”.

Depending on the specific context, not only the interpretation of Ps and Qs but also the relational properties Rs may vary. In the philosophical literature on modeling, the P´s and Q´s are typically interpreted either in ontic or in pragmatic terms. While the ontic view defines Ps and Qs as instances of ontic entities, such as events or components, and understands the property R causally and constitutively (and also, if necessary, formally), the pragmatic view emphasizes the tasks and purposes of modeling. Under the pragmatic interpretation, some elements of models are selected as relevant (or expressed on some continuum of relevance) from a number of available existing elements, because they are (or are seen) relevant to the purpose for which a model is used.

The pragmatic interpretation of relevance can be understood in different ways. Firstly, it can be understood as task based relevance, in which the relevance of P is dependent on

its suitability for task T. It is important to notice that this "suitability" is not dependent on our subjective ways of evaluating that suitability. In short, when relevance is interpreted as task based relevance, then P can be used for the purpose Z (such as explanation), if P is suitable for it (whether or not the user realizes, or even evaluates the relevance of P for it). Secondly, relevance can be understood as intentionality based relevance. On this view, the relevance of P for a certain task Z is understood as dependent on the model user´s (intentional) evaluation of the usability of P for certain purposes. These evaluations are subjective in the sense that the evaluation is viewed as requiring an intentional stance.

However, as I argue in "On Relevance", the intentionality based account is a problematic view of relevance. Firstly, merely postulating or intending a relevance relation to hold between two items does not create a proper relation between them. Secondly, appealing to intentions may actually serve to complicate the problem of relevance rather than offer a possible solution to it. For these reasons, I propose that solving the problem of relevance may require appealing to a more objective account of relevance.

1.7.5. Towards to an Explanation of Conceptual Change: A Mechanistic Alternative

Conceptual change is one of the most studied fields of science education and psychology of learning, and there are hundreds, if not thousands of studies on this topic. These studies show that a crucial aspect of learning in science involves conceptual change i.e. radically organizing and altering the learner's pre- or misconceptions in addition to adding new knowledge to what is already there.

However, despite the fertility of the field there are still important theoretical issues in conceptual change research on which no clear consensus has emerged. Firstly, there is no agreement on what kinds of changes in belief and concept systems constitute conceptual change, and what kinds of changes do not. Secondly, there is no consensus on what the learning mechanisms of conceptual change are. Thirdly, there is no common explanatory framework or paradigm for explaining conceptual change.

In “Towards to an Explanation of Conceptual Change”, I outline a sketch of a mechanistic explanation of conceptual change. I argue that the explanation of an individual learning episode as an instance of conceptual change requires 1) a precise description of the information processing task and 2) a sufficiently accurate and detailed description of the cognitive mechanisms responsible for fulfilling that task. However, although many kinds of cognitive mechanisms of conceptual change have been suggested, sometimes these suggested mechanisms are not mechanisms at all. Moreover, there are very few accounts of conceptual change in which the mechanisms of conceptual change are specified in sufficient detail. These descriptions of “mechanisms” include filler terms, and they fail to satisfy the requirements for genuine mechanism descriptions.

References

Bechtel, W. 2008. *Mental mechanisms: Philosophical perspectives on cognitive neuroscience*. London: Routledge University Press.

Bechtel, W. 2011. Mechanism and biological explanation. *Philosophy of Science*, 78, 533-557.

Bechtel, W. 2012. Understanding endogenously active mechanisms: A scientific and philosophical challenge. *European Journal for the Philosophy of Science*, 2, 233-248

Bechtel, W., & Abrahamsen, A. 2005. Explanation: A Mechanistic Alternative. *Studies in the History and Philosophy of Biomedical Sciences*, 36, 421-441.

Bechtel, W. and Abrahamsen (2010). Dynamic mechanistic explanation: Computational modeling of circadian rhythms as an exemplar for cognitive science. *Studies in History and Philosophy of Science* Part A.,1, 321-333

Bechtel, W., & Richardson, R. 1993. *Discovering Complexity, Decomposition and Localization as Strategies in Scientific Research*. New Jersey: Princeton Univ

Chater, N., Tenenbaum, J. B., & Yuille, A. 2006. Probabilistic Models of Cognition: Conceptual Foundations. *Trends in Cognitive Sciences*, 10, 287-291.

Chater, N., & Vitanyi, P. M. B. 2003. The generalized universal law of generalization. *Journal of Mathematical Psychology*, 47: 346-369.

Callender, C. and Cohen, J. 2006. There Is No Special Problem About Scientific Representation. *Theoria* 55:7-19.

Carruthers, P. 2003. On Fodor's Problem. *Mind and Language*, 18 (5): 502–523.

Colombo, M., Hartmann, S. & van Iersel, R. (forthcoming). Models, Mechanisms, and Coherence. To appear in *The British Journal for the Philosophy of Science*.

Contessa, G. 2011. "Scientific Models and Representation." in *The Continuum Companion to the Philosophy of Science*, ed Steven French & Juha Saatsi. Continuum Press.

Craver, C.F. 2001. Role functions, Mechanisms and Hierarchy. *Philosophy of Science*, 68, 53-74.

Craver, C.F. 2006. When Mechanistic Models Explain. *Synthese,* 153: 355-376.

Craver, C.F. 2007a. *Explaining the Brain: What a Science of the Mind-Brain could be.* New York: Oxford University Press.

Craver, C.F. 2007b. "Constitutive Explanatory Relevance." *Journal of Philosophical Research*, 32:3-20.

Cummins, R. 1983. *The Nature Of Psychological Explanation*. Cambridge, MA.: MIT Press

Cummins, R. 1989. *Meaning and Mental Representation.* Cambridge, MA: MIT Press.

Cummins, R. 2000. "How does it work?" versus "what are the laws?": Two conceptions of psychological explanation. In F. Keil & R. Wilson (Eds.), *Explanation and cognition* (pp. 117-144). Cambridge, MA: MIT Press

Da Costa, N. and French, S. 2000. Models, Theories and Structures: Thirty Years On. *Philosophy of Science* 67: 116-127.

French, S. 2003. A model-theoretic account of representation (or, I don't know much about art…but I know it involves isomorphism). *Philosophy of Science*, 70: 1472–1483.

Frigg, R. 2006. Scientific Representation and the Semantic View of Theories. *Theoria*, 55: 49-65.

Glennan, S. 2002. Rethinking Mechanistic Explanation. *Philosophy of Science* 69: 342-353.

Glennan, S. 2005. Modeling Mechanisms. *Studies in the History and Philosophy of Biomedical Sciences*: 443-464.

Hitchcock, C. 1995. Discussion: Salmon on Explanatory Relevance. *Philosophy of Science* 62:304–20.

Illari, P. & Williamson, J. 2011. Mechanisms are Real and Local. in Illari, P., Russo, F. & J. Williamson (eds), *Causality in the Sciences*. OUP, pp. 818-44.

Imbert, C. 2013. Relevance, Not Invariance, Explanatoriness, Not Manipulability: Discussion of Woodward's Views on Explanatory Relevance. *Philosophy of Science*, 80: 625-636.

Kaplan, D. 2011. Explanation and description in computational neuroscience. *Synthese,* 183 (3): 339-373

Kaplan, D. and Bechtel, W. 2011. Dynamical Models: An Alternative or Complement to Mechanistic Explanations. *Topics in Cognitive Science*, 3: 438-444.

Kaplan, D. and Craver, C. 2011. The Explanatory Force of Dynamical Models. *Philosophy of Science*, 78 (4): 601-627

Knuuttila, T. & Kuorikoski, J. 2011. Idealized Representations, Inferential Devices and Cross-Disciplinary Tools: Theoretical Models in Social Sciences, in Jarvie, I. & Zamora Bonilla, J. (eds.) *The Sage Handbook of the Philosophy of Social Science*. Bodmin: Sage, 530-550.

Kuorikoski, J. 2009. Two concepts of mechanism: componential causal system and abstract form of interaction. International Studies in the *Philosophy of Science,* 23 (2): 143-160

Kuorikoski, J. & Ylikoski, P. 2013. How Organization Explains. In Karakostas, V. & Dieks, D. (eds.) *EPSA11 Perspectives and Foundational Problems in Philosophy of Science*. Springer, 69-80

Lappi, O. & Rusanen, A-M. 2011. Turing Machines and Causal Mechanisms in Cognitive Sciences, In P. McKay Illari, F. Russo & J. Williamson, (eds.) *Causality in the Sciences*. Oxford: Oxford University Press. 2011: 224-239

Machamer, P. K., Darden, L., & Craver, C. 2000. Thinking About Mechanisms. *Philosophy of Science,* 67: 1-25.

Marr, D. 1982. *Vision: A Computational Investigation into the Human Representation of Visual Information*. San Francisco: W.H. Freeman.

Milkowski, M. 2012. Limits of Computational Explanation of Cognition. In Müller, V. (ed.) *Philosophy and Theory of Artificial Intelligence*. Springer.

Milkowski, M. 2013.*Explaining the Computational Mind*. MIT Press.

Mäki, U. 2009. MISSing the World. Models as Isolations and Credible Surrogate Systems. *Erkenntnis* 70:29–43.

Newell, A. 1980. Physical Symbol Systems. *Cognitive Science,* 4(2): 135-183.

Piccinini, G. 2004a. Functionalism, Computationalism and Mental Contents. *Canadian Journal of Philosophy*, 34, 375-410.

Piccinini, G. 2006a. Computational Explanation and Mechanistic Explanation of Mind. In M. DeCaro, F. Ferretti & M. Marraffa (Eds.), *Cartographies of the Mind: The Interface Between Philosophy and Cognitive Science*. Dordrecht: Kluwer.

Piccinini, G. 2006b. Computational Explanation in Neuroscience. *Synthese*, 153, 343-353.

Piccinini, G. 2007. Computing Mechanisms. *Philosophy of Science*, 74: 501-526.

Piccinini, G. 2011. Computationalism, in *Oxford Handbook of Philosophy of Cognitive Science*, Eric Margolis, Richard Samuels, and Stephen Stich, eds., Oxford: Oxford University Press 2011: 222-249.

Piccinini, G. & Craver, C. 2011. Integrating psychology and neuroscience: functional analyses as mechanism sketches. *Synthese*, 183 (3):283-311.

Pöyhönen, S. 2013. *Chasing Phenomena. Studies on Classification and Conceptual Change in the Social and Behavioral Sciences*. Philosophical Studies from the University of Helsinki, Turku: Juvenes Print.

Revonsuo A. 2006. *Inner Presence. Consciousness as a Biological Phenomenon*. Cambridge, MA: MIT Press

Rusanen, A-M. (Under Review). On Relevance. Submitted to EJPS.

Rusanen, A-M. & Lappi, O. 2007. The Limits of Mechanistic Explanation in Cognitive Science. In Vosniadou, S., Kayser, D. & A. Protopapas, (Eds) *Proceedings of the European Cognitive Science Conference 2007*. Howe: Lawrence Erlbaum Associates. 2007: 284-289.

Rusanen, A-M. & Lappi, O. 2012a. An Information Semantic Account of Scientific Models. In H. De Regt, S. Hartmann, & S. Okasha, (eds) *EPSA Philosophy of Science: Amsterdam 2009*, Dordrecht, Springer, 2012: p.315-328.

Rusanen, A-M. & Lappi, O. 2012b. Modeling Cognition: How Fiction Relates to Fact. In N. Miyake, D. Peebles & R. Cooper (Eds) *Building Bridges Across Cognitive Sciences: Proceedings of the 34th Annual Meeting of the Cognitive Science Society*, Austin, TX, 2012: 941-946

Ryder, D. 2004. SINBAD Neurosemantics: A Theory of Mental Representation. *Mind & Language*, 19:211-241.

Sarasevic, T. 2007. Relevance: A Review of the Literature and a Framework for Thinking on the Notion in Information Science. Part II: Nature and Manifestations of Relevance. *Journal of the American Society for Information Science and Technology,* 58(13):1915–1933.

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

Shagrir, O. 2001. Content, Computation and Externalism. *Mind* 110, 369-400.

Shagrir, O. 2006. Why We View the Brain as a Computer. *Synthese*, 153(3): 393-416.

Shagrir, O. 2010. Brains as Analog-Model Computers. *Studies in the History and Philosophy of Science*, 41(3): 271-279.

Shapiro, L. 1997. A Clearer Vision. *Philosophy of Science*, 64, 131-153.

Spohn, W. 2013. A Ranking-Theoretic Approach to Conditionals. *Cognitive Science*, 37: 1074–1106. doi: 10.1111/cogs.12057

Suárez, M. 2003. Scientific Representation: Against Similarity and Isomorphism. *International Studies in the Philosophy of Science*, 17:225-244.

Suárez, M. 2004. An inferential account of Scientific Representation. *Philosophy of Science*, 71:767-779.

Suárez, M. & Solé, A. 2006. On the Analogy between Cognitive Representation and Truth. *Theoria*, 55 (2006): 39-48.

Sun, R. 2008. Theoretical status of computational cognitive modeling, *Cognitive Systems Research*, 2008, doi:10.1016/j.cogsys.2008.07.002

Teller, P. 2001. Twilight of the Perfect Model Model. *Erkenntnits*, 55:393-415.

Weisberg, M. 2007. Who is a Modeler?, *British Journal for Philosophy of Science*, 58, 207–233.

Weiskopf, D. 2011. Models and mechanisms in psychological explanation, *Synthese*, 183: 313-38.

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

Wright, C. & Bechtel, W. 2007. Mechanisms and Psychological Explanation. In P. Thagard (Ed.), *Philosophy of Psychology and Cognitive science* (pp.31-77). Amsterdam: Elsevier

Ylikoski, P. 2013.Causal and Constitutive Explanation Compared. *Erkenntnis*,78(2): 277-297

Ylikoski, P. & Aydinonat, E. (2014).Understanding with theoretical models. *Journal of Economic Methodology*, 21 (1):19-36

Ylikoski, P. & Hedström, P. 2010. Causal Mechanisms in the Social Sciences. *Annual Review of Sociology*, 36: 49-67

Ylikoski, P. K. & Kuorikoski, J. 2010. Dissecting explanatory power. *Philosophical Studies. An international journal for philosophy in the analytic tradition*. 148: 201-219

Ylikoski, P. and Kuorikoski, J. (Under Review). Understanding with External Representations.